MIS Quarterly

RESEARCH ARTICLE

UNIFYING ALGORITHMIC AND THEORETICAL PERSPECTIVES: EMOTIONS IN ONLINE REVIEWS AND SALES[1]

Yifan Yu
Department of Information Systems and Operations Management, Michael G. Foster School of Business, University of Washington, Seattle, WA, U.S.A. {yifanyu@uw.edu}

Yang Yang
Department of Business Management, School of Management, University of Science and Technology of China Hefei, Anhui, CHINA {yy921006@mail.ustc.edu.cn}

Jinghua Huang
Research Center for Contemporary Management, School of Economics and Management, Tsinghua University, Beijing, CHINA {huangjh@sem.tsinghua.edu.cn}

Yong Tan
Department of Information Systems and Operations Management, Michael G. Foster School of Business, University of Washington, Seattle, WA, U.S.A. {ytan@uw.edu}

Emotion artificial intelligence, the algorithm that recognizes and interprets various human emotions beyond valence (positive and negative polarity), is still in its infancy but has attracted attention from industry and academia. Based on discrete emotion theory and statistical language modeling, this work proposes an algorithm to enable automatic domain-adaptive emotion lexicon construction and multidimensional emotion detection in texts. Using a large-scale dataset of China's movie market from 2012 to 2018, we constructed and validated a domain-specific emotion lexicon and demonstrated the predictive power of eight discrete emotions (i.e., surprise, joy, anticipation, love, anxiety, sadness, anger, and disgust) in online reviews on box office sales. We found that representing overall emotions through discrete emotions yields higher prediction accuracy than valence or latent emotion variables generated by topic modeling. To understand the source of the predictive power from a theoretical perspective and to test the cross-culture generalizability of our prediction study, we further conducted an experiment in the U.S. movie market based on theories on emotion, judgment, and decision-making. We found that discrete emotions, mediated by perceived processing fluency, significantly affect the perceived review helpfulness, which further influences purchase intention. Our work shows the economic value of emotions in online reviews, generates insight into the mechanism of their effects, and has managerial implications for online review platform design, movie marketing, and cinema operations.

Keywords: Emotion, online reviews, box office, perceived processing fluency, perceived helpfulness, purchase intention

Introduction

With the prevalence of social media and online review platforms that make reviews ubiquitous, reading reviews has become an integral part of consumers' purchase decisions. Our survey of the movie industry shows that 75.2% of consumers will (and another 21.3% may) refer to online reviews before they decide to buy a movie ticket. The economic value of emotions in online reviews has attracted growing attention from the industry. A management perspective emphasizes the managerial salience of understanding and reacting to emotions (e.g., anger, sadness,

[1] Gal Oestreicher-Singer was the accepting senior editor for this paper. Miguel Godinho de Matos served as the associate editor.

surprise) in large-scale online reviews with the advanced technology of artificial intelligence (AI) and big-data analytics, known as "emotion AI" or "affective computing."[2] Indeed, 25% of Fortune Global 500 companies and over 1,400 brands have adopted emotion AI in their marketing research.[3]

Measuring emotions in online reviews and understanding their effects on consumer behavior and business performance, however, poses a theoretical and algorithmic challenge. There are three fundamental theories of emotion, i.e., dimensional emotion theory, discrete emotion theory, and cognitive appraisal theory (Watson & Spence, 2007). Dimensional emotion theory (Barrett & Russell, 1999) focuses on the effect of emotions from dimensions such as *valence* (positive and negative polarity) and *arousal* (the extent to which a person is energized by an experience) (Yin et al., 2017). Discrete emotion theory[4] identifies specific basic emotions and delineates their effects (Plutchik, 1984; Tomkins, 1962). Cognitive appraisal theory argues that emotional experiences result from cognitive appraisals of an event and its situational environment and highlights the effects of specific emotions, which supports discrete emotion theory (Desmet, 2010). Discrete emotion theory and dimensional emotion theory are two competing theories in terms of measuring consumer emotions[5] (Havlena & Holbrook, 1986; Lerner et al., 2004, 2015; Plutchik & Kellerman, 1982; Russell & Mehrabian, 1974). In view of dimensional emotion theory, existing management research on online reviews has focused mainly on "valence analysis" (e.g., Asur & Huberman, 2010; Hennig-Thurau et al., 2015; Rui et al., 2013; Song et al., 2019), which uses supervised machine learning models to classify a review as positive, negative, or neutral.

More recent management research focuses on "discrete emotion analysis," i.e., detecting the intensities of discrete emotions from large-scale online content (Felbermayr & Nanopoulos, 2016; Malik & Hussain, 2017; Nguyen et al., 2020) by detecting emotion words in texts with existing emotion lexicons. An emotion lexicon contains emotion words and their associated emotional intensities of different emotion categories. State-of-the-art practice, however, fails to consider domain differences and the evolutional nature of emotional expressions[6] (Oliveira et al., 2016; Xue et al., 2014; Yin et al., 2014). Notably, from an algorithmic and data-driven perspective, there is a third emotion analysis approach, which we refer to as "emotion topic analysis," i.e., the use of a topic modeling approach to extract latent emotion topics (Yu et al., 2012).

The motivation for our work originates on the algorithmic side. First, among the three emotion analysis approaches, we consider which is the best to represent emotional information in reviews and of the highest predictive power in regard to sales. We expect valence analysis, the common practice of understanding emotion in online reviews, to be less effective than discrete emotions analysis because, from a theoretical perspective, valence is merely one of the important dimensions of human emotions (Lerner et al., 2015), and lab experiments show that valence-based predictions of consumer cognition and behavioral intention yield contradictory results (Lerner et al., 2004; Yin et al., 2014). We do not know ex ante, however, whether data-driven latent emotion topics would outperform theory-driven discrete emotions. Second, we consider whether combining different representations of emotions produces better predictive power. Investigating these two questions would improve the state-of-the-art algorithmic practice of representing emotions in online content and provide empirical evidence for relevant emotion theories. Third, we consider how to incorporate domain differences and the evolutional nature of emotional expressions and advance the state-of-the-art discrete emotion analysis approach.

To this end, we focus on the movie industry for the following reasons. First, the movie market is of high economic importance, with global box office revenues reaching a record $42.5 billion in 2019.[7] Second, social media and

[2] https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained

[3] https://www.affectiva.com/who/about-us/

[4] Discrete emotions are rooted in human evolution, with expression and recognition fundamentally the same across all individuals, regardless of ethnic or cultural differences (Plutchik, 2001). The theory originated in the late 19th century from Darwin's theory of evolution, which maintains that certain basic emotions evolve from natural selection. Neuroimaging analyses have shown that these discrete emotions are linked to discrete neural signatures and certain structures of the human brain (Saarimäki et al., 2016; Vytal & Hamann, 2010).

[5] On the one hand, Havlena and Holbrook (1986) surveyed 20 participants about their emotions in consumption experiences and found that the emotional dimensions of Russell and Mehrabian (1974) yield good predictions for seven of the eight discrete emotions described in Plutchik and Kellerman (1982). In this regard, emotional dimensions are more favorable than discrete emotions because, without losing much information, less data need to be collected when using emotional dimensions. On the other hand, more recent lab experiments have shown that discrete emotions such as anxiety and anger (Yin et al., 2014) and sadness and disgust (Lerner et al., 2004), are similar in terms of emotional dimensions but have disparate effects on consumers' cognition and purchase intentions, which suggests that the effects of discrete emotions cannot be fully predicted by emotional dimensions.

[6] Emotional expressions are domain dependent and evolutionary over time (Oliveira et al., 2016; Xue et al., 2014) and cannot be fully captured by a general and static emotion lexicon (e.g., RenCECps, LIWC). For example, Yin et al. (2014) used a domain-independent lexicon (LIWC) to capture anger and anxiety words in Yahoo! retailer reviews. Only 2.57% of reviews were detected as containing anxiety and anger words, which was "lower than expected." The authors noted that "these low values are not surprising given the use of a predefined dictionary that does not take context into consideration" (p. 550).

[7] https://www.boxofficepro.com/global-box-office-2019-record-42-5-billion/

online review platform users have considerable interest in discussing movies, with substantial variance in their emotions. Third, the real-world business outcome (i.e., box office revenue) is publicly observable. More importantly, the movie industry was among the most disrupted industries during the COVID-19 pandemic, with billions of dollars lost in the global box office in 2020.[8] As more movie marketing and sales activities go digital, it is timely to leverage the economic value embedded in online user-generated content, which could potentially contribute to the digital resilience of the industry.

In Study 1, from China's movie market, the world's largest, we collected a large-scale dataset that contains 499 movies released between 2012 and 2018 and 3,257,871 microblogging messages related to these movies. We identified 1,214,310 movie reviews from microblogging messages by utilizing a support vector machine (SVM) model and then conducting three emotion analyses on the reviews. First, we conducted valence analysis by following the state-of-the-art implementation in the literature (Hennig-Thurau et al., 2015; Song et al., 2019) and classified reviews as positive or negative. Second, to enable domain-adaptive discrete emotion analysis, we constructed a domain-specific and up-to-date emotion lexicon by extending an existing general emotion lexicon built in 2008, Ren-CECps (Quan & Ren, 2010), and combining it with a popular neural network language model, the Word2Vec model.[9] After validating the newly generated lexicon with out-of-sample testing and human annotation, we utilized it to detect the intensities of eight discrete emotions in reviews, i.e., surprise, joy, anticipation, love, anxiety, sadness, anger, and disgust. We focused on these emotions not only because they are considered the most basic emotions that constitute other emotions, according to discrete emotion theory (Lerner et al., 2004, 2015; Plutchik & Kellerman, 1982; Tomkins, 1962; Yin et al., 2014), but also because they are the most commonly seen in online content (Quan & Ren, 2010). Third, we conducted emotion topic analysis by advancing the approach proposed by Yu et al. (2012). We interpreted the latent emotion topics by leveraging the estimated topic-word distribution and the emotional intensities associated with the emotion words in the lexicon. Based on the results, we infer that there are two prevalent emotions in movie reviews. The first one is positive and activated, resembling "excited" or "delighted"; the second one is less positive and of low arousal, resembling "bored" or "disappointed."

After obtaining three representations of emotions, and combining them with other control predictors, we used an expanding window strategy (Geva et al., 2017; Song et al., 2019) to predict box office sales with four machine learning models, i.e., linear regression (LR), random forest (RF), support vector regression (SVR), and XGBoost (XGB), which are either popular or state of the art for numeric prediction tasks.

We found that first, compared to a basic lexicon, our newly constructed domain-specific and up-to-date lexicon for discrete emotion analysis achieved significantly higher prediction accuracy for box office sales across all prediction models, which demonstrates the practical value of addressing the domain difference and evolutional nature of emotional expressions. Second, discrete emotions achieved significantly higher predictive power than valence. Under the SVR model, all discrete emotions were able to *individually* achieve a higher prediction accuracy than valence could. In addition, discrete emotions were of significantly higher predictive power than latent emotion topics. Third, combining discrete emotions with valence or latent emotion topics did not improve prediction accuracy, which implies that the emotional information carried by valence and latent emotion topics was absorbed in discrete emotions. These findings demonstrate the economic value of discrete emotions in online reviews and provide empirical evidence for discrete emotion theory.

The algorithmic study further motivated us to investigate the source of the predictive power of discrete emotions in online reviews from a theoretical perspective. If the algorithmic predictive relationship is causal with an interpretable mechanism, the robustness of our predictive study can be further justified, and one would expect the predictive results to be generalizable to other contexts in which a similar mechanism exists. Moreover, although the existing work on emotion in online reviews has examined the effects of anxiety and anger on perceived helpfulness (Yin et al., 2014) and happiness and anger on product evaluation (Kim & Gupta, 2012; Xiao et al., 2018), the effect sizes of other discrete emotions remain understudied. As emotional experiences consist of eight discrete emotions (Plutchik & Kellerman, 1982), the effects of emotions in reviews cannot be fully understood until we have knowledge of a comprehensive set of discrete emotions. More importantly, although the existing mechanisms, i.e., perceived writer cognitive effort (Yin et al., 2014) and perceived writer rationality (Xiao et al., 2018), help us understand the effects of anxiety and anger, we lack a mechanism to understand the effects of other discrete emotions. Finally, we were motivated by the potentially cross-

[8] https://en.wikipedia.org/wiki/Impact_of_the_COVID-19_pandemic_on_cinema

[9] The Word2Vec model is a commonly used statistical language model proposed by Mikolov et al. (2013). It maps a word to a high-dimensional vector that indicates the meaning of the word in a context, using deep-learning techniques. It can exploit the semantic information of words and judge the semantic similarity between words by calculating the cosine distance between corresponding word vectors. Compared with other statistical language models, such as the neural network language model and recurrent neural network language model, Word2Vec provides better performance on measuring semantic word similarities, with a much lower computational cost (Mikolov et al., 2013).

cultural differences in the effects of emotions (Eid & Diener, 2001), which made us consider whether the effects of emotions in online reviews remain significant for English users.

We thus conducted a randomized experiment in Study 2 to investigate the causal impacts of emotions in online reviews on consumer purchase intention and their underlying mechanism in the U.S. movie market. We found that all positive (negative) discrete emotions had a significantly positive (negative) effect on purchase intention. Discrete emotions had higher effect sizes than do the important factors investigated by the previous studies, which highlights their economic significance. Second, mediated by *perceived processing fluency* (the extent of feeling the ease of cognitive processing) (Reber et al., 2002), discrete emotions significantly affected the perceived review helpfulness. Such a mediation effect was significant when subjects were exposed to all discrete emotions, except anger. Further, the effect of anger was completely mediated by perceived writer rationality. Finally, we confirmed that perceived helpfulness significantly influences purchase intention. The remainder of this article is organized as follows. First, we summarize the literature on the economic significance of emotions in reviews and propose a theoretical framework to explain the effects. Second, we elaborate on the analyses and results of Studies 1 and 2. Finally, we discuss the contribution and managerial implications of this work.

Literature Review and Theory

Economic Significance of Emotions in Online Reviews

Emotions affect cognitive outcomes, including judgment and decision-making (Lerner et al., 2015; Loewenstein, 2000). Emotions in reviews affect consumers' perceived review helpfulness[10] and their attitudes toward products (Kim & Gupta, 2012; Yin et al., 2014). Table 1 provides a summary of the key factors related to perceived review helpfulness, product evaluation, purchase probability, and purchase intention. Anger, anxiety, valence, and arousal (East et al., 2017; Kim & Gupta, 2012; Xiao et al., 2018; Yin et al., 2014, 2016, 2017) are of comparable or higher effect sizes in terms of other factors, e.g., review quality, credibility (Teng et al., 2017), and reviewer characteristics (Huang et al., 2015; Ngo-Ye & Sinha, 2014; Yin et al., 2016), showing the economic significance of these emotions in online reviews. Despite their potential significance, the effects of discrete emotions other than anger and anxiety are understudied, which is an important concern because we cannot expect the effects of anger and anxiety to be generalized to other discrete emotions. According to the cognitive appraisal theory of emotion (Lerner et al., 2004, 2015) and neuroimaging research on emotion (Saarimäki et al., 2016; Vytal & Hamann, 2010), discrete emotions are independent of each other not only because they are evoked by different cognitive appraisals but also because they are linked to discrete neural signatures and certain structures of the human brain. Further, the existing mechanisms, i.e., perceived writer cognitive effort (Yin et al., 2014) and perceived writer rationality (Xiao et al., 2018), are built on anger and anxiety. The mechanism to explain the effects of other discrete emotions in online reviews, however, remains understudied.

Mechanism of Emotions in Online Reviews on Purchase Intention

From a theoretical perspective, we establish the link between emotions in online reviews and purchase intention. Specifically, we argue that mediated by perceived processing fluency, emotions affect consumers' perceived helpfulness of reviews, and perceived helpfulness further influences purchase intention. First, emotions in reviews affect perceived processing fluency. According to the "feelings as information" framework (Schwarz, 1990; Shiv, 2007), emotional information, serving as information sources (Isbell et al., 2013) or as information cues (DeSteno et al., 2000), influences cognitive processes, including consumers' perceptions, intentions, and behavior (So et al., 2015). Further, affective priming theory (Klauer, 1997) argues that when a reader is primed by emotional information, the fluency of subsequent information processing is affected.[11] Perceived processing fluency emerges at the perceptual level when a reader feels the ease of grasping a physical identity of a stimulus (Reber et al., 2002). Further, psychological research has empirically shown that emotions can positively or negatively affect the fluency of information processing (Zeelenberg et al., 2006).

[10] Perceived helpfulness is a cognitive result of reviews as well as a predictor of product sales (Mudambi & Schuff, 2010). Perceived helpfulness of reviews is defined as the extent to which the reviews are perceived by consumers to facilitate their purchase-decision process (Yin et al., 2014). When reviews are perceived as helpful, the recall of these reviews will affect consumers' attitudes and intentions (Filieri et al., 2018).

[11] The "expectancy" mechanism can explain the effect: Compared to nonemotional information, an evaluative response to emotional information is activated more quickly and automatically; then, based on the evaluation of emotional information, the reader forms an expectancy toward the nonemotional information, which interferes or augments the evaluative response to the nonemotional information, depending on whether the expectancy is congruent with the evaluation of the nonemotional information (Klauer, 1997).

Table 1. Summary of Key Factors

Dependent variable	Independent variable	Normalized effective size[a]	Mechanisms tested	Source
Perceived helpfulness	Valence (rating)	Not significant	NA	Mudambi & Schuff (2010)
		0.07	NA	Schindler & Bickart (2012)
		0.07	NA	Huang et al. (2015)
		NA	Confirmation bias	Yin et al. (2016)
	Anger	0.11	Perceived cognitive effort	Yin et al. (2014)
	Anxiety	0.36	Perceived cognitive effort	Yin et al. (2014)
	Reviewer engagement	NA	NA	Ngo-Ye & Sinha (2014)
	Reviewer impact	NA	NA	Huang et al. (2015)
	Review ranking	0.12	NA	Yin et al. (2016)
	Review length	0.15	NA	Yin et al. (2016)
	Arousal	0.18	Perceived effort	Yin et al. (2017)
	Adjectives, state verbs, action verbs, readability, subjectivity	NA	NA	Krishnamoorthy (2015)
Product evaluation	Anger	-0.38	Review information value	Kim & Gupta (2012)
		-0.25	Perceived problem seriousness, perceived reviewer rationality	Xiao et al. (2018)
	Valence	0.24	NA	Teng et al. (2017)
	Review quality	0.17	Emotional strength	Teng et al. (2017)
Purchase probability	Positive/negative word of mouth	NA	NA	East et al. (2017)
Purchase intention	Review credibility	0.20	NA	Teng et al. (2017)
	Two-sided reviews	0.25	Information helpfulness	Filieri et al. (2018)
	Food-related disgust	-0.16	NA	Shimp & Stuart (2004)

Note: [a]The effective sizes are normalized as the number of units changing in the dependent variable when there is a one-unit change in the independent variable. NA = not applicable.

Consumers who actively search for online reviews usually have a positive attitude toward a product (East et al., 2017). Due to confirmation bias (i.e., the tendency to accept information in reviews that confirm one's initial beliefs) (Yin et al., 2016), consumers can more easily process and accept review information that contains positive discrete emotions. Reviews with negative discrete emotions stand in contrast to consumers' prior attitudes toward the product. Consumers tend to search reviews for diagnostic information that they have not considered (Hennig-Thurau et al., 2015), which requires a higher level of cognitive effort (Yin et al., 2014). Thus, we expect positive (negative) discrete emotions to increase (decrease) perceived processing fluency.

Second, perceived processing fluency improves perceived review helpfulness.[12] A review presented in fluent information is perceived as more helpful than one in disfluent information (Chen & Sakamoto, 2016) because processing fluency increases the credibility of online reviews, contributing to perceived helpfulness (Li et al., 2013; Tsekouras, 2017). Recent studies have shown that processing fluency facilitated by an easy-to-read font (Chen & Zhang, 2018), review rankings (Huang et al., 2014), review ratings (Risselada et al., 2018), and language style matching (Liu et al., 2019) helps improve perceived review helpfulness. Based on existing empirical evidence, we expect that emotion-induced perceived processing

[12] This is a theoretically interesting relationship because, formally, helpfulness concerns *what* information is contained in the review, and whether the information is considered *diagnostic* by the consumer in the process of making purchase decisions (Yin et al., 2014), which is linked to the theory of information diagnosticity (Hennig-Thurau et al., 2015). Processing fluency concerns *how* the diagnostic information, if any, is conveyed (more specifically, in an easy- or hard-to-process way). A rational individual should be able to determine the helpfulness of a review, which is independent of how the information is conveyed, and independent of whether it is easy or hard to process, as long as it is *understandable*. There are likely some interesting behavioral mechanisms that establish the link, for example, limited attention, cognitive capacity, and bounded rationality (Egeth & Kahneman, 1975). Indeed, management researchers are increasingly focusing on this interesting topic (Chen & Sakamoto, 2016; Chen & Zhang, 2018; Liu et al., 2019).

fluency will improve perceived review helpfulness.[13] Third, it is well established in the literature that perceived review helpfulness has a positive effect on purchase intention (e.g., Filieri et al., 2018). Finally, we do not rule out the mechanisms in the existing literature by which emotions in online reviews affect perceived review helpfulness and purchase intention (see "Mechanisms Tested" in Table 1), but we choose to focus on perceived processing fluency, which differentiates this paper from the existing literature. Figure 1 provides a visual summary of the theoretical framework.

Study 1: A Field Study on Prediction

Predicting future sales is a critical issue in various business domains. We focus on the Chinese movie market, the world's largest. In this industry, however, the movie theater's operation efficiency is surprisingly low. Among 9,504 theaters, only 15 achieve over 50 million RMB of revenue per year, and the average occupancy rate is only around 15%.[14] In what follows, we explain how weekly and daily box office predictions can enhance theater operation efficiency.

First, weekly box office predictions contribute to movie selection decisions. In this market, most theaters are small, with an average of only 5.3 screens.[15] Nevertheless, the total number of new movies in one month could be 30 to 40, which means that theater managers should choose one from at least three new movies for a single screen. These decisions are usually made on a weekly basis. An improper movie screening decision could lead to a 0.21 million RMB loss in revenue per week.[16] Thus, optimizing movie selection decisions according to box office sales predictions is of great importance for a majority of theaters. Second, daily box office predictions help daily screening room arrangements. Screening rooms in theaters are usually of different sizes (ranging from tens to hundreds of seats). If one movie is predicted to be popular the next day, the cinema managers can move it to a larger room or temporarily add more screening schedules to avoid selling out of tickets.

Third, daily demand predictions contribute to cinemas' inventory management. Cinemas offer low-profit-margin tickets to attract audiences and sell high-profit-margin food and drinks to make a profit. The inventory management decisions regarding perishable food and drinks need to be made on a daily basis. Thus, each day, cinema managers face a newsvendor problem, whereby they need to order a certain amount of perishable inventory in advance to meet the next day's uncertain demand. Accurate predictions of daily demand are thus critical to reducing the cost of inventory. To address these managerial challenges, our work first focuses on daily box office predictions and includes additional analysis of weekly predictions.

Data Collection

We invited 14 research assistants and collected all publicly available movie samples from 2012 to 2018 from maoyan.com, a public movie database in China. For each movie, we collected daily box office revenues from maoyan.com and all microblogging reviews related to the movie from weibo.com, the largest microblogging platform in China, for the duration of its screening period. We deleted movies without any reviews, leaving us with a sample of 589 movies. In addition, we deleted movie samples that contained missing data in their box office revenue, leaving us with 499 movies in our sample. The number of movie and date combinations was 12,993 and the number of microblogging messages was 3,257,871. The average total box office revenue for a movie was 286.5 million RMB. The average box office revenue for a movie per day was 11.2 million RMB, and the average number of a movie's screening days was 16.9.

To the best of our knowledge, the scale of our empirical test is the largest among the extant studies on microblogging messages and movie box office sales, which usually involve fewer than 100 movies or have less than a one-year observation period (Hennig-Thurau et al., 2015; Liu, 2006; Rui et al., 2013; Song et al., 2019; Yu et al., 2012). This sample mimics a practitioner's situation of wanting to use all of the available data to achieve the best predictive accuracy. In addition, we downloaded the natural language processing and information retrieval (NLPIR) microblog corpus (over 5 million microblogging messages). We combined the NLPIR corpus and the collected 3.26 million movie-related microblogging messages as training data for the Word2Vec model.

[13] Our focus is related to but different from the above recent studies because we leverage the idea of *perceived* processing fluency from advertising research (Storme et al., 2015). As we detail later, the presentation of the information (the contrast of texts, fonts, and the cognitive content of the review) in our experiment is the same for the treatment and the control group, except that the review in the treatment group is manipulated by adding emotional expressions. Thus, the processing fluency of reviews is affected by emotional expressions at the *perceptual* level, consistent with affective priming theory (Klauer, 1997).

[14] http://tech.sina.com.cn/i/2018-06-24/doc-iheirxye9028421.shtml

[15] http://finance.eastmoney.com/news/1355,20180110820566856.html

[16] https://www.douban.com/group/topic/50452206/

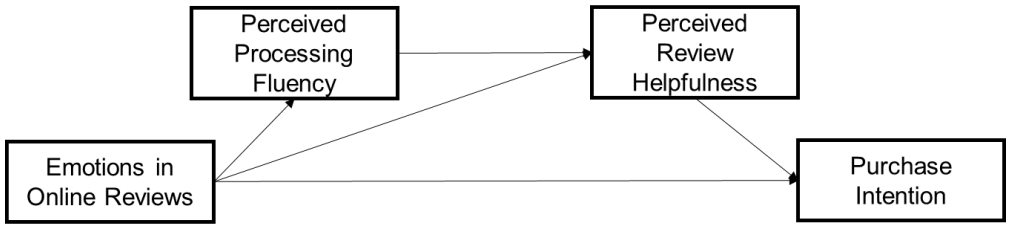
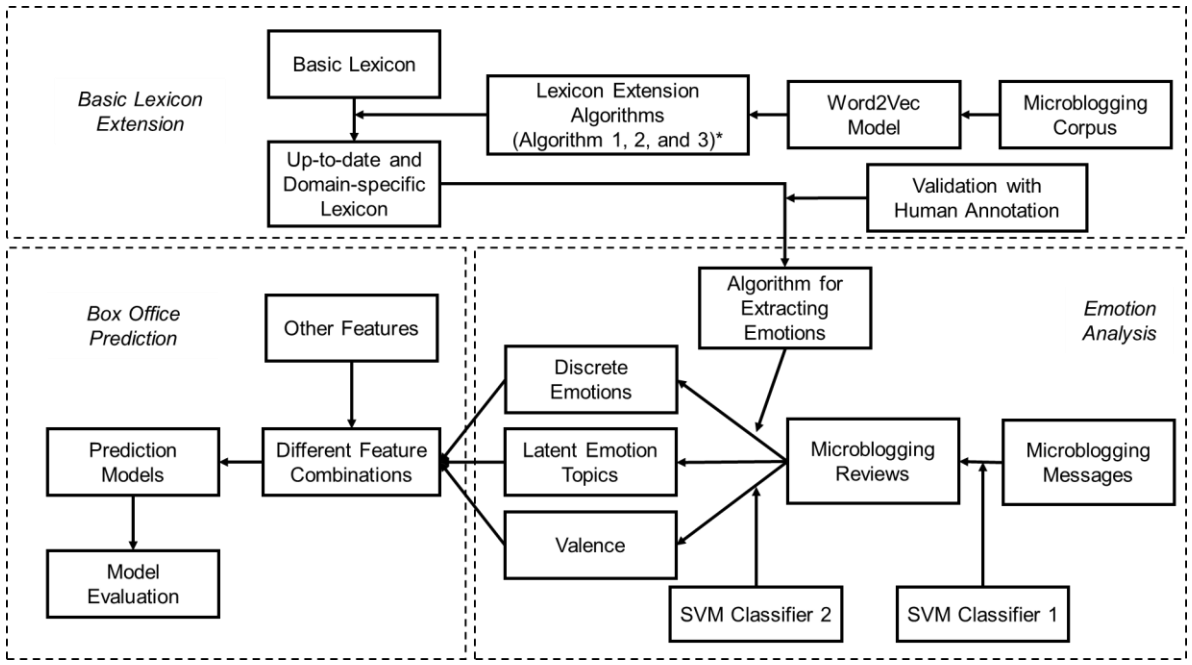


Figure 1. Theoretical Framework



Note: *Algorithms 1, 2, and 3 are detailed in the following section.

Figure 2. Empirical Framework

Algorithms and Empirical Analyses

We present our empirical research framework in Figure 2. Our method involves three parts: basic lexicon extension, emotion analysis, and box office sales predictions. First, we trained the Word2Vec model, extending the basic lexicon into an up-to-date and domain-specific emotion lexicon. Second, we conducted emotion analysis of microblogging reviews. To obtain microblogging reviews, we dropped nonreview messages with a pretrained SVM classifier. Then, we conducted valence analysis, discrete emotion analysis, and latent emotion topic analysis of the reviews. Third, we combined emotion features along with other features, including historical box office sales, microblogging review volume, and screening days and weekends, to generate box office sales predictions and investigate the predictive power of different emotion features.

Basic Lexicon Extension

Our approach requires an existing emotion lexicon as a basic lexicon. Ren-CECps is an emotion lexicon based on 1,487 Chinese blog texts and contains 22,406 unique emotion words (Quan & Ren, 2010). We filtered out 7,179 words that did not appear in any of the collected microblogging messages, leaving us with a lexicon that contains 15,227 words. Ren-CECps focuses on eight discrete emotions, i.e., *surprise*, *joy*, *anticipation*, *love*, *anxiety*, *sadness*, *anger*, and *disgust*. Each word w_i in Ren-CECps is associated with an eight-dimension

emotion-intensity vector $v_i = \{e_j^i\}_{j=1}^{N=8}$, where $e_j^i \in [0,1]$ is the manually annotated intensity of the j^{th} discrete emotion.

We selected Ren-CECps for the following three reasons. First, these eight discrete emotions are the most commonly expressed emotions in Chinese blog texts; using these emotions decreases confusion in emotion-category selection (Quan & Ren, 2010). Second, the texts are annotated based on Chinese blogs. Blogs and microblogging messages are online user-generated content. Hence, we believe that Ren-CECps is more suitable for our context than lexicons based on general Chinese texts (e.g., NTUSD, HowNet, DUT). Finally, Ren-CECps is manually annotated and statistically validated (Quan & Ren, 2010).[17] Compared to the algorithm-constructed emotion lexicons (e.g., Yang et al., 2016), the results of Ren-CECps are more precise and reliable.

Next, we trained a Word2Vec model by using 3.26 million Chinese movie-related microblogging messages from the microblogging platform and 5 million microblogging messages from the NLPIR microblog corpus (Appendix A). The Word2Vec model maps each word in the training text to a high-dimensional vector, termed a *word vector*, which contains the semantic information of the word (Mikolov et al., 2013). The Word2Vec model defines the similarity between two words as the cosine similarity between their associated vectors (Mikolov et al., 2013; Song et al., 2018).[18]

Algorithm 1 presents our lexicon extension algorithm. First, we randomly split the basic lexicon $L_0 = \{(w_i, v_i)\}_{i=1}^{N}$ into training (80%), validation (10%), and test (10%) sets, i.e., L_0^{tr}, L_0^{va}, and L_0^{te}. Second, we constructed W as a set of all unique words in the target domain, except stop words. From W, we constructed a subset P that contains potential emotion words.

Specifically, for each word $w_i \in W$, if $w_i \notin L_0^{tr}$, we obtained M words that are most similar to w_i: $\{w_{i,j}\}_{j=1}^{M}$ by using the pretrained Word2Vec model. If $\exists j\, s.t.\, w_{i,j} \in L_0^{tr}$, $w_{i,j}$ is considered a potential emotion word (Xue et al., 2014), and we add $w_{i,j}$ to P. Third, we iteratively extended the basic lexicon L_0^{tr}. We defined algorithm $f(w; L, \theta)$ that maps a word w to an emotion-intensity vector $v_i = (e_1, e_2, \ldots, e_8)$, where $e_i \in [0,1]$. As we detail in Algorithm 2, f is conditional on parameter θ and lexicon L. For each word $w_i \in P$, we used f to map w_i to an emotion-intensity vector v_i. If v_i has at least one positive intensity value, we added a word-intensity pair (w_i, v_i) to the basic lexicon. After one iteration, we checked whether the number of words in the basic lexicon increased. If so, we repeated the iteration until the number of words converged.

Next, in Algorithm 2, we illustrate how we map a word to emotional intensities (f). The rationale of f originated from the *K*-nearest neighbor algorithm, but we customized it in our task and introduced hyperparameter $\alpha \in [0,1]$ to control for noise. In particular, for word w, we constructed set S_w that contains K-most similar words by using the Word2Vec model. We then checked the intersection of S_w and emotion lexicon L, denoted as S_w^L. If S_w^L was non-empty, we averaged the emotion intensities of emotion words in S_w^L to determine the emotion intensities of w. We compressed emotion intensities that were lower than α to zero to reduce noise.

We used validation set L_0^{va} to optimally choose parameters K and α in f by using a random parameter search (Bergstra & Bengio, 2012) and then used test set L_0^{te} to evaluate the out-of-sample performance of f, as shown in Algorithm 3. We first drew a set Θ of $\theta = (K, \alpha)$, where K is randomly selected from $\{1,2, \ldots, 10\}$ and $\alpha \sim uniform(0, 0.2)$. For each θ, we used $f(w; L_0^{tr}, \theta)$ to generate a predicted emotion-intensity vector v for w. Then, we used the mean absolute error (MAE) to evaluate the error between v and the human-annotated value v^0. After we obtained the optimal parameter θ^* that minimizes the validation MAE, we evaluated the out-of-sample MAE of $f(w; L_0^{tr}, \theta^*)$ based on test set L_0^{te}. The optimal parameter $\theta^*$is (K^*, α^*) = (5, 0.15), and the corresponding validation MAE is 0.072 (for full parameter-search results, see Appendix B). The MAE for the test set is 0.073. Given that the MAE can vary from 0.0 to 1.0, this result demonstrates both the in-sample and out-of-sample validity of our algorithm. Alternatively, our algorithm f can be constructed by using the mean model[19] or directly training more sophisticated state-of-the-art models, i.e., RF and XGB, on the 200-dimensional word vectors. As we detail in Appendix C, however, additional analysis shows that although RF and XGB can outperform the mean model, our algorithm achieves the best performance.

[17] Eleven annotators participated in the annotation work. According to Quan and Ren (2010), the authors spent two months on the joint training of annotators and developed annotation instructions. They also used a Kappa statistic to measure the pairwise agreement among the 11 annotators. The Kappa coefficient of the agreement is a statistic adopted by the computational linguistics community as a standard measure for such a purpose. The agreement for emotional words is 0.785. Given the complexity of this annotation task, we believe that the annotations are reliable and valid.

[18] The natural language processing module *Gensim* in Python is used to construct the model. The dimension of a word vector is set to 200. This parameter is a default setting suggested by Gensim and used in the state-of-the-art implementation of Chinese word embeddings, which are validated across different Chinese natural language processing tasks (Song et al., 2018).

[19] As a benchmark model, the prediction of a new word in the mean model is produced by using the average emotion intensities of all words in the training set, i.e., $f_{mean}(w; L_0^{tr}) = \sum_i v_i / |L_0^{tr}|$, where $v_i \in L_0^{tr}$.

Algorithm 1. Lexicon Extension	
Pseudocode	**Remarks**
Input a basic lexicon $L_0 = \{(w_i, v_i)\}_{i=1}^{N}$;	w_i is an emotion word and v_i is an eight-dimension vector that represents emotional intensities of w_i.
Randomly split L_0 into training (80%), validation (10%), and test (10%) sets: L_0^{tr}, L_0^{va}, and L_0^{te};	
Get all words from corpus W;	W is the set of all unique words, except stopping words, in a target domain.
Input a pre-trained Word2Vec model;	
FOR each word $w_i \in (W - L_0^{tr})$:	
Get M words that are most similar to w_i: $\{w_{i,j}\}_{j=1}^{M}$;	M is selected as 100 (Xue et al., 2014).
IF $\exists j\ s.t.\ w_{i,j} \in L_0^{tr}$:	
Add $w_{i,j}$ into the Potential Emotion Word Set P;	
END IF	
END FOR	
Initiate lexicon $L_{old} = L_0^{tr}$, $L_{new} = L_0^{tr}$;	L_{new} will be the output as an extended lexicon.
Input an algorithm $f(w; L, \theta)$, a mapping from a word w to emotion intensities $(e_1, e_2, \ldots, e_8)$, where $e_i \in [0,1]$;	f is defined conditional on a parameter θ and a lexicon L. The details of function f are provided in Algorithm 2.
WHILE TRUE:	An iterative process
FOR each word $w_i \in P$:	
$v_i \leftarrow f(w_i; L_{old}, \theta) = (e_1, e_2, \ldots, e_8)$;	
IF $\sum_i e_i > 0$:	If a word has at least one positive intensity value, add it to the extended lexicon.
$L_{new} \leftarrow L_{new} \cup \{(w_i, v_i)\}$;	
END IF	
END FOR	
IF $\lvert L_{new} \rvert = \lvert L_{old} \rvert$:	Check convergence
BREAK	
ELSE:	
$L_{old} \leftarrow L_{new}$;	Update the basic lexicon and further extend the lexicon.
RETURN L_{new};	

Algorithm 2. Mapping Word to Emotional Intensities (Implementing f)	
Pseudocode	**Remarks**
Input a lexicon L;	L can be, for example, L_0^{tr} in Algorithm 1.
Input parameter $\theta = (K, \alpha)$; $K \in Z^+, \alpha \in [0,1]$;	Z^+ represents the set of all positive integers.
Initiate outcome $v = (e_1, e_2, \ldots, e_8)$; $\forall i, e_i \leftarrow 0$;	
Input a pre-trained Word2Vec model;	
Input word w;	
Get a set of K most similar words to w in the Word2Vec model: $S_w \leftarrow \{w_j\}_{j=1}^{K}$;	
$S_w^L = S_w \cap L$;	Find all of the most-similar words that are in the lexicon.
IF $\lvert S_w^L \rvert = 0$:	If the interaction is an empty set, return an all-zero
RETURN $v = (e_1, e_2, \ldots, e_8)$;	vector and end the algorithm.
END IF	
FOR each word $w_j \in S_w^L$:	
$d_j \leftarrow$ semantic similarity between w_j and w;	The similarity is measured by the cosine similarity between the word vectors of w_j and w (Mikolov et al., 2013).
$v_j = (e_1^j, e_2^j, \ldots, e_8^j) \leftarrow$ retrieve emotional intensities of w_j from L;	

END FOR	
FOR $i \in \{1,2,\ldots,8\}$:	
$e_i \leftarrow \frac{\sum_{j=1} d_j e_i^j}{\sum_j d_j}$;	Estimate the emotional intensities of w by the weighted average of its most similar words.
IF $e_i < \alpha$: $e_i \leftarrow 0$;	If the magnitude is less than a threshold, the value will be compressed to zero to reduce noise.
END IF	
END FOR	
RETURN $v = (e_1, e_2, \ldots, e_8)$;	

Algorithm 3. Validate and Test f	
Pseudocode	**Remarks**
Input a basic lexicon $L_0 = \{(w_i, v_i)\}_i^N$;	L_0 is the same as that in Algorithm 1.
Randomly split L_0 into training (80%), validation (10%), and test (10%) sets: L_0^{tr}, L_0^{va}, and L_0^{te};	
Draw a set Θ of $\theta = (K, \alpha)$, where K is randomly selected from $\{1,2,\ldots,10\}$ and $\alpha \sim uniform(0, 0.2)$;	Random parameter search
FOR each $\theta \in \Theta$:	Find the best hyperparameters with the validation set.
FOR each $w_i \in L_0^{va}$:	
$v_i^0 = (e_1^0, e_2^0, \ldots, e_8^0)$ ←retrieve true values of emotional intensities of w_i from the validation set;	
$v_i \leftarrow f(w_i; L_0^{tr}, \theta) = (e_1, e_2, \ldots, e_8)$;	Get predicted values with the training set.
Get MAE $\epsilon_i \leftarrow \sum_{i=1}^{8} \frac{\lvert e_i - e_i^0 \rvert}{8}$;	
END FOR	
Calculate average validation MAE $\epsilon_\theta \leftarrow \sum_i \frac{\epsilon_i}{\lvert L_0^{va} \rvert}$;	
END FOR	
Find $\theta^* \leftarrow \underset{\theta \in \Theta}{\operatorname{argmin}}\, \epsilon_\theta$;	
FOR each $w_i \in L_0^{te}$:	Calculate test errors with the test set.
$v_i^0 = (e_1^0, e_2^0, \ldots, e_8^0)$ ←retrieve true values of emotional intensities of w_i from the test set;	
$v_i \leftarrow f(w_i; L_0^{tr}, \theta^*) = (e_1, e_2, \ldots, e_8)$;	
Get MAE $\epsilon_i \leftarrow \sum_{i=1}^{8} \frac{\lvert e_i - e_i^0 \rvert}{8}$;	
END FOR	
Calculate the average test MAE $\epsilon_{\theta^*} \leftarrow \Sigma_i \frac{\epsilon_i}{\lvert L_0^{te} \rvert}$;	
RETURN ϵ_θ and ϵ_{θ^*};	

Next, we used Algorithm 1 and f obtained in Algorithms 2 and 3 to extend the basic lexicon. To maximize the model accuracy, we combined the training and validation sets as a basic lexicon. We used the test set to evaluate the out-of-sample MAE after each iteration in Algorithm 1. As shown in Figure 3, Algorithm 1 stopped after 11 iterations, and a total of 6,710 new emotion words were obtained. This shows the value of lexicon extension, without which at least 30.6% (6,710 out of 21,937) of the emotional expressions would not be captured. During the iterations, the test MAEs were consistently between 0.067 and 0.068, demonstrating the validity of the iteration process.

To further ensure the validity of the newly mined emotion words, we randomly selected 1,000 newly mined words and followed the manual annotation procedure by Quan and Ren (2010) to evaluate the accuracy of the estimated emotion intensities of these words. Because the emotional expression of a word depends on its context, for each word, we randomly retrieved 10 different movie reviews that contained the word. Then, we recruited and trained ten research assistants to annotate the emotional intensities of these words based on their contexts. For each word w_i ($i \in \{1,2,\ldots,1000\}$) in each review r_{ij} ($j \in \{1,2,\ldots,10\}$), two research assistants independently annotate the intensities of the eight discrete emotions. We averaged the two assistants' annotation results to ensure that the results were not biased toward either research assistant's subjectivity. We denoted the averaged annotation as $v_{ij} = \left(e_{ij}^1, e_{ij}^2, \ldots, e_{ij}^8\right)$, where $e_{ij}^k \in [0,1]$ represents the intensity for the kth discrete emotion.

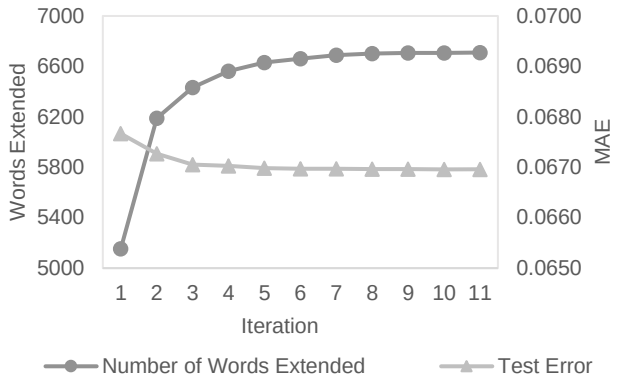


Figure 3. Number of Words Extended and Test Errors

Next, we averaged over all the context j of word w_i and obtained $\overline{v_i} = \left(\frac{\sum_j e_{ij}^1}{10}, \frac{\sum_j e_{ij}^2}{10}, \ldots, \frac{\sum_j e_{ij}^8}{10}\right) \triangleq (e_i^1, e_i^2, \ldots, e_i^8)$ as ground truth emotional intensities for w_i. Finally, we denoted the estimated emotional intensities for w_i as $\widehat{v_i} = (\hat{e}_i^1, \hat{e}_i^2, \ldots, \hat{e}_i^8)$ and calculated the mean absolute error as $MAE = \frac{\sum_i \sum_k \left|e_i^k - \hat{e}_i^k\right|}{1000 \times 8}$. We determined that the MAE is 0.080, which is close to the out-of-sample testing error of 0.073, demonstrating the validity of the newly mined emotion words. Table 2 provides examples of emotion words and their estimated emotional intensities to further illustrate the performance of our algorithm. First, our algorithm correctly distinguishes the subtle differences between similar words; it predicts that "admire" is used to express love, whereas "awe" is a blend of love and anxiety. "Grief" has a higher level of sadness than "sad." "Shocking" is predicted to have a higher level of surprise than "pleasantly surprised," whereas the latter is a blend of surprise and joy. Second, our algorithm can detect domain-specific words. For example, the algorithm detects the word "photogenic," a word specific to the movie review domain, and predicts that it can, to some extent, express love.

Emotion Analysis

Valence analysis: We conducted a two-stage data processing approach to determine the valence of microblogging messages (Hennig-Thurau et al., 2015; Song et al., 2019). First, we conducted content classification to identify 1,214,310 microblogging review messages from 3,257,871 microblogging messages and filtered out nonreview messages, such as spam and unrelated messages. In particular, four coders, who were extensively trained for this annotation task, manually coded 20,000 randomly selected microblogging messages into review or nonreview messages. Every microblogging message was individually labeled by two different coders. 16,244 messages received consistent annotations across different coders, 7,193 of which were coded as reviews. We randomly divided the 16,244 messages into a training set, validation set, and test set. With the training set (80% of the annotated 16,244 messages), we conducted Chinese word segmentation using a commonly used tool, Jieba, built in Python. We leveraged the feature hashing approach and the chi-square feature selection method to obtain 5,000 features (Song et al., 2019) and train an SVM model based on the selected features. With the validation set (10% of the annotated 16,244 messages), the best parameters were selected as an RBF-kernel with a penalty coefficient of C = 300 by random parameter searching. Using the test set (another 10% of the annotated 16,244 messages), we determined that the out-of-sample F1-score is 86.3%. Second, the two coders divided 7,193 microblogging review messages into positive and negative messages,[20] and we followed the same procedure to obtain another RBF-kernel SVM model, with an optimal C equal to 2,500 and an out-of-sample F1-score equal to 90.8% (for accuracy, 94.8%).

[20] We used a binary-score annotation instead of a continuous-score annotation for valence because it is the standard procedure of the literature on online review sentiment analysis (Hennig-Thurau et al., 2015; Rui et al., 2013; Song et al., 2019). It helps to build a baseline for us to understand the extent to which our approach can outperform the standard procedure. Further, compared to a binary-score annotation, a continuous-score annotation for valence could lead to noise in the annotated results, as it is difficult to achieve agreement among different annotators when they are rating in a continuous manner (Pang & Lee, 2005). Finally, although it is binary at the review level, it is continuous at the daily or weekly level at which we conduct prediction. This allows valence to vary continuously in our prediction model.

Table 2. Examples of Emotion Words and Estimated Emotional Intensities

Word in Chinese	Word in English	Surprise	Sadness	Love	Joy	Disgust	Anticipation	Anxiety	Anger
敬佩	Admire	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00
敬畏	Awe	0.00	0.00	0.24	0.00	0.00	0.00	0.15	0.00
悲伤	Sad	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.00
悲痛	Grief	0.00	0.81	0.00	0.00	0.00	0.00	0.00	0.00
惊喜	Pleasantly surprised	0.21	0.00	0.00	0.16	0.00	0.00	0.00	0.00
震惊	Shocking	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00
镜头感	Photogenic	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00

Note: The translation is produced by Google Translate (translate.google.com).

We then sorted 1,214,310 reviews into positive and negative reviews using the SVM model. We denoted the number of positive reviews for movie m on day t as $Pos_{m,t}$ and total reviews as $ReviewVol_{m,t}$. Then, the review valence was calculated as the ratio of positive reviews to total reviews, i.e., $ReviewVal_{m,t} = \frac{Pos_{m,t}}{ReviewVol_{m,t}}$ (Hennig-Thurau et al., 2015). The mean and standard deviations for $ReviewVol_{m,t}$ are 93.5 and 164.6, respectively. The mean and standard deviation for $ReviewVal_{m,t}$ are 0.8 and 0.2, respectively.

Discrete emotion analysis: Following the document-level emotion expression space model (Quan & Ren, 2010), we mapped each review r_j into an eight-dimensional vector $\{E_k(r_j)\}_{k=1}^{N=8}$, where $E_k(r_j)$ is determined by emotion, negation, and degree words contained in review r_j, representing the intensity of the kth discrete emotion. Negation and degree words are frequently used in Chinese and are thus helpful for accurate emotion analysis (Quan & Ren, 2010). We adopted a negation-word dictionary provided by TextMind, a Chinese language psychological analysis system developed by the Chinese Academy of Sciences. The dictionary contains 31 frequently used negation Chinese words. We adopted a degree-word dictionary containing sixty degree words and their annotated values from Shi (2017). For example, the degree value of the word meaning "the most" in Chinese was annotated as 1.5, and that of the word meaning "kind of" was annotated as 0.8. We used a forward-sliding window to capture these negations and degree words. If our algorithm found an emotion word in the review, it checked the three words[21] before the emotion word to capture negation and degree words. Therefore, we have:

$$E_k(r_j) = \frac{1}{n_k}\sum_{i=1}^{n_k}(-1)^{m_i} \times DegV_i \times I_k(w_i), \quad (1)$$

where $\{w_i\}_{i=1}^{n_k}$ are the emotion words[22] in both review r_j and our lexicon and express the k^{th} discrete emotion. If $n_k = 0$, then $E_k(r_j)$ is set to zero. $I_k(w_i)$ refers to the k^{th} discrete emotion intensity of w_i. m_i is the total number of negative words that appear in the sliding window of w_i. $DegV_i$ is the average degree value of all degree words that appear in the sliding window of w_i. Let $v_{m,t}$ denote the total number of reviews on day t for movie m. Let $e_{m,t,k}$ represent the kth discrete emotion intensity of movie m on day t, and let $r_{m,t,i}$ denote the ith review on movie m on day t. Thus, we have:

$$e_{m,t,k} = \frac{\sum_{i=1}^{v_{m,t}} E_k(r_{m,t,i})}{v_{m,t}}, k \in \{1,2,\ldots,8\}. \quad (2)$$

Latent emotion topic analysis: We are motivated by Yu et al. (2012), who adopted a topic modeling approach (latent semantic indexing [LSI]) to capture latent emotion topics and use them for sales predictions. We followed their

[21] We chose 3 as the sliding-window size because, in Chinese, negative and degree words usually appear within three words prior to the corresponding emotion word.

[22] One might be concerned that emotion words in movie reviews could be used not only to subjectively express emotions but also to describe the movie plot. The latter could induce noise in understanding consumer emotion when using a lexicon-based approach. We thus conducted an additional study to manually check the extent to which emotion words are plot-description related or subjective-expression related. In particular, we randomly selected 200 movie reviews and invited two research assistants to read the reviews, identify all the emotion words in the reviews, and sort the emotion words into a plot-description-related class or a subjective-expression-related class. There are 607 emotion words identified from the reviews, 544 (89.6%) of which are identified as subjective-expression-related emotion words. This result indicates that a majority of the emotion words in online movie reviews are used to express one's subjective feelings toward movies and justifies that it is appropriate to adopt a lexicon-based approach to analyze movie viewers' emotions. Further, we note that even plot-description-related emotion words can be helpful in predicting movie sales. For example, these emotion words signal the genre information of a movie. Such information is related to box office revenue and may be useful in a machine learning model.

methodological framework but used a more advanced topic modeling approach, latent Dirichlet allocation (LDA). The LDA approach can overcome the overfitting issue in LSI and provide better latent topic presentations (Blei et al., 2003). We followed the framework by Yu et al. (2012) to formalize the problem: For a given set of reviews $R = \{r_1, r_2, \dots, r_N\}$ and a set of emotion words from lexicon $L = \{w_1, w_2, \dots, w_M\}$, the review data can be described as an $N \times M$ matrix $C = \{c(r_i, w_j)\}_{ij}$, where $c(r_i, w_j)$ is the word frequency of w_j in r_i. As such, each row in C is an emotion-word frequency vector that corresponds to a review. We assume that each observation (r_i, w_j), the emotion expression in the review, is generated by an unobservable latent emotion $Z \in \{z_1, z_2, \dots, z_k\}$.

Similar to LDA, for which hidden factors represent the "topics" of the document (Blei et al., 2003), Z represents the type of latent emotion that the writer would like to express through word w_j (Yu et al., 2012). Then, an LDA with the online variational Bayes algorithm is implemented with the Python machine learning tool sklearn to estimate the probability that latent emotion Z is embedded in review r, i.e., $P(Z|r)$ and the probability that the emotion word w would be used when the latent emotion is Z, i.e., $P(w|Z)$. Then, each review r_i is mapped into a k-dimensional latent emotion vector LaE_i, where $LaE_i = \left(P(z_1|r_i), P(z_2|r_i), \dots, P(z_k|r_i)\right)^T$, which is the probability distribution of latent emotions embedded in r_i. Similar to Equation (2), we can summarize LaE_i at the daily level. We use $LaE_{m,t,k}$ to denote the k^{th} latent emotion topic of microblogging review messages on movie m on day t. A key parameter to select is the number of latent emotions, i.e., k. Perplexity, defined as $e^{-\bar{v}}$, where $\bar{v}$ is the log-likelihood per word, is commonly used to select this parameter (Blei et al., 2003). The lower the perplexity that an LDA model produces, the better the fitness of the model to the data will be. Figure 4 shows that the perplexity is minimized when $k = 2$, i.e., the optimal number of latent emotions is 2.[23] The latent emotion estimated by Yu et al. (2012) is not interpretable. In our work, however, we conducted a deeper analysis and derived some insight into these two latent emotions. Formally, given a latent emotion z_k $(k = 1,2)$, the expected intensity of a discrete emotion e_s $(s = 1,2, \dots, 8)$ can be derived as $E(e_s|z_k) = \sum_{w \in L} E(e_s|w) \cdot P(w|z_k)$, where $E(e_s|w)$ is given by lexicon L, and $P(w|z_k)$ is from the LDA model. We present the results of $E(e_s|z_k)$ in Figure 5.

One might expect that the two latent emotions would simply be "positive" or "negative," but we found that the message sent by $E(e_s|z_k)$ involved more than these valences. First, we found that z_2 is more positive than z_1 because, for positive discrete emotions (love and joy), z_2 has an intensity that is about two times as high as z_1, whereas, for negative discrete emotions (anger, anxiety, disgust, and sadness), the differences between z_1 and z_2 are much smaller. Further, we observed that for each discrete emotion s, $E(e_s|z_1)$ is less than $E(e_s|z_2)$, which means that to express latent emotion z_2, a writer would use emotion words that are of high emotional intensity. In other words, z_2 is of higher arousal than z_1. Thus, based on the results, we infer that the two prevalent emotions in movie reviews are (1) positive and activated, with strong love and joy, which resembles "excited" or "delighted," and (2) less positive and of low arousal, which resembles "bored" or "disappointed."

Box Office Sales Predictions

We adopted a monthly expanding window approach for box office sales predictions, which has been commonly used by earlier predictive research (Geva et al., 2017; Song et al., 2019). We chose such an approach because it mimics a practitioner's situation, whereby every month, the cinema manager collects all of the available historical data to train a prediction model. In particular, we assumed that the time point that a manager adopts the approach was January 1, 2018. Then, for each month t after January 1, 2018, the data samples of the preceding months (month 1 to $t-1$, where month 1 is January 2012) were used as a training set and the data samples of month t were used as the test set. Our lexicon was created by using only the reviews before January 1, 2018. As such, information about patterns of the future was not used to "predict" the past data points. We used historical daily box office revenue data, emotions in reviews, review volume, screening days, and weekends as predictive variables (Table 3), consistent with the existing box office prediction literature (Song et al., 2019; Yu et al., 2012).

The extant research has noted that many movie-specific time-invariant factors could be influential for total box office sales, including genre, prerelease buzz, production budget, and star and director effects (Hennig-Thurau et al., 2015). With a time-series approach, the effects of these factors can be absorbed in historical box office variables.

[23] We clarify that perplexity is minimized based only on the training data. In particular, we first used the lexicon constructed by using only the reviews before January 1, 2018 to capture the emotion words that appear in all of the reviews. Then, we used emotion word frequency only in the reviews before January 1, 2018 to train the LDA model and select the number of topics. Then, for reviews after January 1, 2018, we used the LDA model to predict the latent emotion topics of the reviews. Finally, the predicted latent emotion topics were used as features to predict the movie sales performance after January 1, 2018. This mimics a practitioner's situation, whereby the practitioner trains the LDA and prediction model with all of the historical data available at a given time point and then uses trained models to predict the outcomes after this time point.

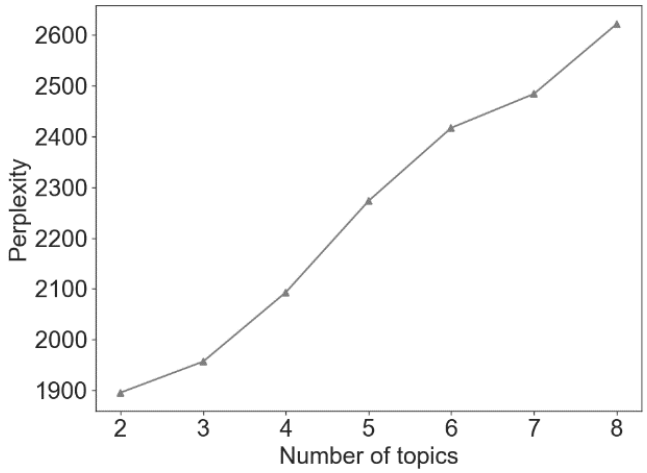


Figure 4. Perplexity of the LDA Model across the Number of Topics

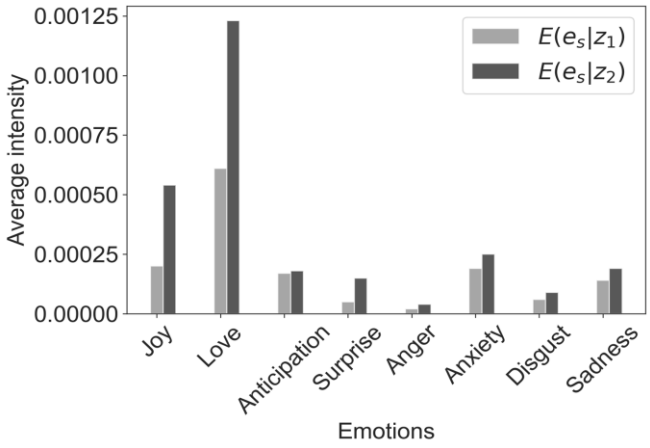


Figure 5. Average Discrete Emotion Intensities in Latent Emotions

Table 3. Description of Variables

Variable	Description
$BoxOffice_{m,t}$	The box office of movie *m* on day *t* (RMB yuan)
$ReviewVol_{m,t}$	Microblogging review message volume of movie *m* on day *t*
$ReviewVal_{m,t}$	Microblogging review message valence of movie *m* on day *t*
$e_{m,t,k}$	The *k*th ($k \in \{1,2,\ldots,8\}$) discrete emotion intensity of microblogging review messages on movie *m* on day *t*
$LaE_{m,t,k}$	The *k*th ($k \in \{1,2\}$) latent emotion topic of microblogging review messages on movie *m* on day *t*
$Age_{m,t}$	Total number of movie *m* screening days from its release day to day *t*, controlling for the tendency of the box office
$IsWeekend_{m,t}$	Dummy variable for indicating whether day *t* is the weekend for movie *m*, controlling for the seasonality of the box office induced by weekends

To compare the predictive power of different features, we defined feature sets, as presented in Table 4. The parameter p is the maximum number of time lags. To avoid generating too many undetermined parameters in our prediction model, we used the same time lag effect p for historical box office sales, review volume, valence, and emotion variables. We built prediction models with four different algorithms, i.e., LR, RF, SVR, and XGB, with default settings recommended by sklearn. For a given feature set, we first used a default RF model to conduct feature selection on a training set and then used the selected features to predict box office sales with one of the four algorithms.

Finally, we evaluated prediction model performance using the mean absolute percentage error (MAPE) between the predicted and observed box office sales on any given day, consistent with the literature (Song et al., 2019; Yu et al., 2012). Specifically, MAPE = $\frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t-F_t}{A_t}\right|$, where A_t is the actual sales of a movie-date pair $t \in \{1,2,\dots,n\}$ $(n = 929)$ in the test set, and F_t is the predicted sales. We used the LR model and base feature set to determine time lag p. Among $p \in \{1,2,\dots,7\}$, the MAPE is minimized when $p = 3$. Thus, we used $p = 3$ for further analysis.

Results

Descriptive Results

We present the correlational analysis of discrete emotions and other predictors in Table 5. All of the absolute correlations between discrete emotions and other box office predictors are no more than 0.21. The low correlations show the orthogonality of emotions and other predictors, indicating that emotions provide important complementary information for the existing predictors. We also make several observations. First, all emotions are slightly positively correlated with review volume. This is perhaps because more-popular movies would also have a higher volume of reviews, and more-popular movies are often better at evoking audience emotions. Second, the four negative emotions, i.e., anger, anxiety, disgust, and sadness, are slightly negatively correlated with review valence. The two positive emotions, i.e., love and joy, are slightly positively correlated with valence. Surprise and anticipation are considered mixed emotions in the existing literature (Nguyen et al., 2020) and can be either positive or negative in valence. For example, a negative review that contains surprise could be: "This movie is surprisingly bad." The results show that on average, surprise and anticipation are slightly negatively correlated with valence. Third, all emotions, except anticipation, are slightly positively correlated with the weekend timing of a movie. This is not unexpected because (1) on weekends, there are typically larger audiences and more reviews, and (2) review volume is positively correlated with emotions. In addition, the slightly negative correlation of anticipation may be due to people expressing their expectations about movie watching on the upcoming weekend on weekdays. Finally, all emotions are slightly negatively correlated with the number of screening days since a movie was released ($Age_{m,t}$). If viewers are enthusiastic about a movie, they are likely to watch the movie immediately after its release. Later adopters are typically less emotionally engaged with the movie than earlier adopters, as the results suggest. All of these observations are consistent with common experience and thus show the validity of the discrete emotion variables. We present the correlations between each pair of the eight emotions in Table 6. No correlation is no more than 0.31, which validates the orthogonality across discrete emotions and justifies the necessity of investigating all of the emotions.

Model Predictions

To demonstrate the value of constructing an up-to-date and domain-specific lexicon, we compared the prediction performance when using the extended and basic lexicons. First, we constructed an up-to-date and domain-specific lexicon based on the reviews from 2012 to 2017 (Lexicon 2017). We constructed another domain-specific lexicon based only on reviews up to 2012 (Lexicon 2012). Second, we used the basic lexicon (built in 2008), Lexicon 2012, and Lexicon 2017 to analyze eight types of discrete emotional features in the reviews of 2018. The sets of eight discrete emotions generated by the basic lexicon and Lexicon 2012 are denoted as emo8_basic and emo8_2012, respectively (emo8 is generated by Lexicon 2017). Tables 7a and 7b present the model prediction results. Rows (a.0) and (b.0) show the MAPE of two benchmark feature sets, i.e., emo8 and base+emo8 (that combine the base feature set and emo8; in Tables 7a and 7b, the plus sign indicates the combining of different feature sets), under four prediction models. The lowest MAPE (39.99) is achieved by base+emo8 under SVR.[24]

[24] Predicting movie sales is a difficult task. This MAPE has largely outperformed prior research, which has primarily used online review sentiments for box office predictions (Yu et al., 2012) (in the same setup with the time lag as 3, the prior research's MAPE is over 80). We believe that with more data available, such as advertising expenditure data, practitioners could further reduce the MAPE.

Table 4. Feature Sets and Variables

Feature sets	Variables
Base	Historical daily box office data, historical review volume, and control variables: $\{BoxOffice_{m,t-1}, BoxOffice_{m,t-2}, \ldots, BoxOffice_{m,t-p}, ReviewVol_{m,t-1}, ReviewVol_{m,t-2}, \ldots, ReviewVol_{m,t-p}, Age_{m,t}, IsWeekend_{m,t}\}$.
Val	Historical review valence variables: $\{ReviewVal_{m,t-1}, ReviewVal_{m,t-2}, \ldots, ReviewVal_{m,t-p}\}$.
emo8	Historical eight types of discrete emotion variables: $\{e_{m,t-1,k}, e_{m,t-2,k}, \ldots, e_{m,t-p,k}\},\ k \in \{1,2,\ldots,8\}$.
Anger	If we use only one type of discrete emotion as a feature set, the feature set will be named after this kind of emotion. For example, let $k = 1$ denote the emotion "anger"; then, the *Anger* feature set is: $\{e_{m,t-1,1}, e_{m,t-2,1}, \ldots, e_{m,t-p,1}\}$. Similarly, by choosing different values of the parameter k, we can define feature sets *Anxiety*, *Disgust*, *Sadness*, *Anticipation*, *Surprise*, *Love*, and *Joy*.
Latent	Historical latent emotion topic variables: $\{LaE_{m,t-1,k}, LaE_{m,t-2,k}, \ldots, LaE_{m,t-p,k}\}, k \in \{1,2\}$.

Table 5. Correlations between Emotions and Other Box Office Sales Predictors

Correlation	$ReviewVol_{m,t}$	$ReviewVal_{m,t}$	$IsWeekend_{m,t}$	$Age_{m,t}$
$Joy_{m,t}$	0.09	0.00	0.02	-0.10
$Love_{m,t}$	0.12	0.05	0.01	-0.08
$Surprise_{m,t}$	0.09	-0.15	0.02	-0.08
$Anticipation_{m,t}$	0.10	-0.08	-0.00	-0.08
$Anger_{m,t}$	0.03	-0.16	0.02	-0.04
$Anxiety_{m,t}$	0.10	-0.19	0.02	-0.08
$Disgust_{m,t}$	0.14	-0.21	0.01	-0.05
$Sadness_{m,t}$	0.12	-0.08	0.01	-0.07

Note: All correlations are statistically significant ($p < 0.01$).

Table 6. Correlations between Discrete Emotions

Correlation	$Joy_{m,t}$	$Love_{m,t}$	$Surprise_{m,t}$	$Anticipation_{m,t}$	$Anger_{m,t}$	$Anxiety_{m,t}$	$Disgust_{m,t}$	$Sadness_{m,t}$
$Joy_{m,t}$	1.00							
$Love_{m,t}$	0.07	1.00						
$Surprise_{m,t}$	0.02	0.04	1.00					
$Anticipation_{m,t}$	0.03	0.02	0.17	1.00				
$Anger_{m,t}$	0.08	-0.03	0.06	0.03	1.00			
$Anxiety_{m,t}$	0.11	0.15	0.12	0.12	0.09	1.00		
$Disgust_{m,t}$	0.07	0.02	0.10	0.09	0.30	0.15	1.00	
$Sadness_{m,t}$	0.09	0.16	0.10	0.17	0.04	0.31	0.13	1.00

Note: All correlations are statistically significant ($p < 0.01$).

Table 7a. MAPE of Emotion Feature Sets across Different Prediction Models

No.	Feature sets	LR	RF	SVR	XGB	No.	Feature sets	LR	RF	SVR	XGB
a.0	MAPE of emo8	116.76	124.90	118.21	135.42						
MAPE increase compared to emo8						**MAPE increase compared to base + emo8**					
a.1	emo8_basic	11.65*** (3.45)	33.82*** (10.81)	6.82*** (1.72)	25.32** (14.78)	a.12	Sadness	15.28*** (5.74)	186.74*** (61.13)	10.23*** (3.98)	33.26** (17.70)
a.2	emo8_2012	9.39*** (3.29)	18.55** (9.60)	1.40** (0.65)	12.78 (11.05)	a.13	Surprise	15.95*** (5.29)	170.96*** (46.27)	12.09*** (3.87)	81.72*** (27.62)
a.3	Latent	19.74*** (6.33)	193.74*** (62.58)	32.78*** (8.82)	86.73*** (33.93)	a.14	val+Anger	12.65*** (4.66)	106.4*** (33.39)	14.65*** (4.48)	85.03*** (34.82)
a.4	Latent + emo8	4.87*** (1.65)	20.87** (9.01)	8.04*** (2.29)	14.57** (7.92)	a.15	val+Anxiety	14.46*** (5.05)	149.76* (90.84)	16.84*** (5.13)	84.52*** (31.57)
a.5	Val	12.98*** (4.74)	71.72*** (22.57)	18.44*** (5.65)	56.04*** (22.29)	a.16	val+Anticipation	15.89*** (5.35)	59.2*** (20.75)	17.12*** (5.03)	48.19*** (18.38)

a.6	Anger	10.28*** (4.07)	83.61*** (33.73)	4.17** (2.09)	39.40** (18.31)	a.17	val+Disgust	18.15*** (6.27)	72.12** (34.76)	16.21*** (4.99)	47.99** (26.63)
a1.7	Anxiety	11.92*** (4.39)	83.79*** (22.09)	7.83*** (2.81)	36.31*** (13.40)	a.18	val+Joy	13.95*** (5.64)	51.22*** (17.90)	14.95*** (5.20)	76.20*** (31.94)
a.8	Anticipation	14.02*** (4.88)	86.01*** (32.10)	9.82*** (3.33)	37.33** (19.88)	a.19	val+Love	21.38*** (6.16)	69.68*** (27.08)	21.51*** (5.80)	56.73*** (23.62)
a.9	Disgust	14.30*** (5.34)	15.34* (9.98)	7.97*** (2.90)	-14.38 (11.43)	a.20	val+Sadness	15.88*** (5.70)	106.95*** (35.79)	16.50*** (5.35)	33.26** (17.90)
a.10	Joy	14.00** (6.42)	66.95*** (25.41)	9.66** (4.98)	67.27 (55.55)	a.21	val+Surprise	17.30*** (5.82)	74.01*** (21.34)	18.87*** (5.59)	96.66*** (33.07)
a.11	Love	30.48*** (10.11)	154.72*** (61.86)	15.97*** (4.45)	65.40*** (26.34)	a.22	val+emo8	22.02*** (6.64)	72.88*** (29.99)	17.66*** (5.09)	60.71** (26.87)

Note: Standard deviations in parentheses. MAPE = $\frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t-F_t}{A_t}\right|$, where A_t is the actual value of a movie-date pair $t \in \{1,2,\dots,n\}$ $(n = 929)$ in the testing set, and F_t is the predicted value. For Row (a.1) to (a.22), the MAPE increase is derived by $\frac{1}{n}\sum_{t=1}^{n}100\left(\left|\frac{A_t-F_{tj}}{A_t}\right|-\left|\frac{A_t-F_{t0}}{A_t}\right|\right)$, where F_{t0} is the predicted value of emo8, and F_{tj} is that of the feature set in Row (a.j). The standard deviation of the MAPE increase is the standard deviation of $100\left(\left|\frac{A_t-F_{tj}}{A_t}\right|-\left|\frac{A_t-F_{t0}}{A_t}\right|\right)$ divided by $\sqrt{n}$ due to the central limit theorem. Then, a one-sided *t*-test is used to determine the significance level of the MAPE increase. *p < 0.10, **p < 0.05, ***p < 0.01.

Table 7b. MAPE of Emotion and Base Feature Sets across Different Prediction Models

No.	Feature sets	LR	RF	SVR	XGB	No.	Feature sets	LR	RF	SVR	XGB
b.0	MAPE of base+emo8	44.40	44.87	39.99	41.00						
MAPE increase compared to emo8						**MAPE increase compared to base + emo8**					
b.1	base+emo8_basic	0.05 (0.13)	1.22 (1.08)	0.06** (0.04)	1.78* (1.22)	b.13	Base+surprise	1.33** (0.66)	3.12** (1.45)	0.70 (0.67)	0.36 (1.66)
b.2	base+emo8_2012	0.20 (0.19)	0.45 (0.98)	0.03 (0.04)	1.02 (1.04)	b.14	Base+val+anger	1.37** (0.67)	1.40 (1.54)	1.30** (0.68)	0.79 (1.04)
b.3	base+latent	1.35** (0.66)	2.41* (1.58)	1.67*** (0.68)	1.94 (1.68)	b.15	Base+val+anxiety	1.36** (0.64)	4.95*** (1.86)	1.38** (0.69)	1.82 (1.58)
b.4	base+latent+emo8	0.33*** (0.11)	0.00 (1.33)	0.59*** (0.15)	2.31** (1.14)	b.16	Base+val+anticipation	1.32** (0.66)	4.67*** (1.87)	1.34** (0.68)	2.13* (1.48)
b.5	base+val	1.26** (0.67)	5.12*** (1.86)	1.02* (0.72)	0.28 (1.37)	b.17	Base+val+disgust	1.26** (0.64)	4.34*** (1.70)	1.30** (0.69)	1.17 (1.43)
b.6	Base+anger	1.37** (0.66)	1.37 (1.63)	0.68 (0.67)	0.85 (1.26)	b.18	Base+val+joy	1.31** (0.66)	2.30* (1.55)	1.32** (0.69)	2.94** (1.53)
b.7	Base+anxiety	1.34** (0.65)	3.41*** (1.43)	0.72 (0.67)	1.69 (1.72)	b.19	Base+val+love	1.33** (0.67)	1.22 (1.19)	1.27** (0.67)	2.18* (1.45)
b.8	Base+anticipation	1.31** (0.66)	3.68*** (1.34)	0.69 (0.66)	2.98** (1.56)	b.20	Base+val+sadness	1.42** (0.66)	5.23*** (1.79)	1.36** (0.68)	2.05* (1.58)
b.9	Base+disgust	1.24** (0.64)	4.42*** (1.71)	0.69 (0.67)	0.71 (1.25)	b.21	Base+val+surprise	1.29** (0.66)	2.82** (1.51)	1.32** (0.69)	1.48 (1.45)
b.10	Base+joy	1.30** (0.66)	3.13** (1.41)	0.70 (0.67)	3.35** (1.75)	b.22	Base+val+emo8	0.15 (0.14)	0.38 (0.90)	0.40*** (0.10)	2.74*** (1.13)
b.11	Base+love	1.33** (0.66)	4.76*** (1.45)	0.66 (0.65)	0.91 (1.46)	b.23	Base	1.41** (0.74)	8.80*** (2.01)	6.19*** (1.06)	6.54*** (2.07)
b.12	Base+sadness	1.39** (0.66)	4.27*** (1.62)	0.70 (0.67)	1.19 (1.64)						

Note: Standard deviations in parentheses. The same procedure used in Table 7a is applied to Rows (b.1) to (b.23), where F_{t0} indicates the predicted value of base+emo8, and F_{tj} indicates the predicted value of the feature sets in Row ($b.j$). *p < 0.10, **p < 0.05, ***p < 0.01.

For Rows (a.1) to (a.22), the *MAPE increase*, compared to emo8, is derived by $\frac{1}{n}\sum_{t=1}^{n}100\left(\left|\frac{A_t-F_{tj}}{A_t}\right|-\left|\frac{A_t-F_{t0}}{A_t}\right|\right)$, where F_{t0} is the predicted value of emo8; F_{tj} is that of feature set in Row (a.j). A positive MAPE increase in Row (a.j) indicates that using emo8 produces a lower MAPE than using the feature set in Row (a.j). The standard deviation of the MAPE increase (as shown in parentheses in the table) is the standard deviation of $100\left(\left|\frac{A_t-F_{tj}}{A_t}\right|-\left|\frac{A_t-F_{t0}}{A_t}\right|\right)$ for $t \in \{1,2,\dots,n\}$ divided by $\sqrt{n}$ due to the central limit theorem. Then, a one-sided *t*-test was used to determine the significance level of the MAPE increase. The same procedure was applied to Rows (b.1) to (b.23), where F_{t0} indicates the predicted value of base+emo8, and F_{tj} indicates the predicted value of the feature sets in Row (b.j). The feature sets in Row (b.j) are derived by combining the base feature set with the feature set in Row (b.j). We mainly focus on interpreting the results presented in Table 7b, because the base features are expected to be used in practice. Table 7a serves as a benchmark and robustness check for Table 7b.

The MAPE increase in Table 7b is smaller than that in Table 7a. It implies that the predictive advantage that emo8 has over the other emotion feature sets partially results from its association with features in the base feature set (historical sales, review volume, seasonality, and timing). Notably, base+emo8 is clearly the method of choice in our setting because its predictive power is always superior or at least equivalent to that of the other alternative feature sets studied in this work.

Value of Lexicon Extension

Rows (b.1) and (b.2) of Table 7b show that emo8_basic and emo8_2012 consistently generate a higher MAPE than emo8. Rows (a.1) and (a.2) of Table 7a show that these results are qualitatively consistent and statistically more significant. The MAPE differences between emo8 and emo8_2012 result only from constructing an up-to-date lexicon, as they are both generated by domain-specific lexicons. The MAPE differences between emo8 and emo8_2012 are mostly smaller than those between emo8 and emo8_basic, which indicates that in addition to constructing an up-to-date lexicon, constructing a domain-specific lexicon plays an important role in reducing the MAPE. The MAPE differences between base+emo8_2012 and base+emo8 are insignificant in Table 7b, implying that constructing a domain-specific lexicon is more effective in reducing the MAPE than constructing an up-to-date lexicon. In other words, although emotional expressions within the domain of movie reviews evolves from 2012 to 2017, the difference in emotional expressions caused by such an evolution is smaller than the difference caused by domain specificity.

Discrete Emotions and Latent Emotion Topics

We then investigated the predictive power of latent emotions as compared to discrete emotions. Row (b.3) of Table 7b shows that discrete emotions outperform latent emotion topics. Row (b.4) of the table shows that combining discrete emotions and latent emotion topics achieves a worse prediction result than using discrete emotions alone. Rows (a.3) and (a.4) in Table 7a show that these results are qualitatively robust and statistically significant. The results can be explained by discrete emotion theory, which argues that complex emotions are mixtures of discrete emotions. The information of the latent emotions, as complex emotions, should already be absorbed by the discrete emotions; thus, latent emotions bring redundant information to predictions. Combining emo8 and latent emotions may lead to overfitting issues. SVR and XGB are more flexible models and are more likely to suffer from overfitting issues than LR and RF. Accordingly, base+latent+emo8 has a larger MAPE increase under SVR and XGB than under LR and RF in Row (b.4) of Table 7b. Further, latent emotion topics are generated based on the data-driven approach, i.e., topic modeling (Blei et al., 2003; Yu et al., 2012). There is no existing theory, to our knowledge, guaranteeing that they are unbiased measures of consumer emotions. Therefore, measurement errors may contaminate the prediction.

Discrete Emotions and Valence

Next, we investigated the predictive power of valence, discrete emotions, and the combination of the two sets of features. Rows (a.5) and (b.5) show that discrete emotions outperform valence in predictive power. Rows (a.22) and (b.22) of the table show that combining valence and discrete emotions cannot increase the predictive power, compared to using discrete emotions alone, which suggests that discrete emotions and valence are substitutive rather than complementary. Compared to discrete emotions, valence is a relatively simple and biased proxy of consumer sentiment (Lerner et al., 2015). For example, consumers' emotions with similar valence (sadness and disgust, anxiety, and anger) are linked to distinct cognitive and behavioral reactions (Lerner et al., 2004; Yin et al., 2014). Further, if valence is just a biased predictor, the orthogonality of valence and emotions (as shown in Table 5) would not benefit the prediction in a meaningful way. Therefore, similar to latent emotions, the issues of overfitting and measurement errors may affect valence as well. For each discrete emotion, we make similar conclusions. Under SVR (the model with the highest accuracy and lowest variance), Rows (a.5) to (a.21) and Rows (b.5) to (b.21) show that each discrete emotion generates a lower MAPE than they do in combination with valence or when using valence alone. Third, the values of MAPE increase in Rows (a.6) to (a.13) and Rows (b.6) to (b.13) are all positive (except disgust in Row (a.9) under XGB), indicating that combining all discrete emotions (emo8) outperforms using each discrete emotion alone. Indeed, discrete emotion theory (Lerner et al., 2015; Plutchik & Kellerman, 1982; Tomkins, 1962) argues that discrete emotions are relatively independent of each other (consistent with the results in Table 6). To combine all of the emotions means to combine orthogonal and meaningful information. The above exception (disgust) indicates that disgust has the highest predictive power among the eight discrete emotions. This is also consistent with our experimental results, presented later, where we show that the effects of disgust on perceived review helpfulness and purchase intention are greater than the effects of most of the other emotions. Nevertheless, the best performance is achieved by using base+emo8, which indicates other discrete emotions still provide complementary predictive power to disgust.

Additional baseline models: To further establish the relevance of the machine learning approach and all of the machinery involved, we tested two more baseline models against our best model, base+emo8 under SVR. First, we used only the previous-day sales variable as the predictor. We found that the MAPE for the first baseline model is 46.73, which is significantly higher than that of base+emo8 (46.73 vs. 39.99, paired *t*-test *p*-value < 0.001). Second, we used the Word2Vec representation for reviews and combined this with the base feature set to predict next-day sales. Specifically, we mapped each word in a review to a 200-dimension word vector. Thus, the Word2Vec representation for a review is the average of all word vectors. Research has empirically shown that this is a successful and efficient way of obtaining sentence semantic information (Arora et al., 2017; Kenter et al., 2016). It is also a common practice in deep learning to first learn low-level, high-dimension, and latent representations for unstructured textual data and then combine them with the traditional machine learning models, such as SVR, to predict the outcomes. Using this baseline model achieved a statistically significantly lower MAPE than using the base feature set (45.39 vs. 46.18, paired *t*-test *p*-value = 0.093), showing that latent semantic representations provide useful information beyond the base feature set. Nevertheless, the MAPE of this baseline model is still significantly higher than the MAPE when using base+emo8 (45.39 vs. 39.99, paired *t*-test *p*-value < 0.001).[25]

Weekly Predictions

We included an additional analysis of weekly predictions. We summarized our variables to a weekly level and followed the same procedure to make weekly predictions. We found that the MAPE by the base feature set is 57.53%, which is the same as that by combining the Base and Val feature sets. This indicates that the valence feature is not predictive of weekly box office sales and is dropped by the RF model in the feature selection process. Combining discrete emotions and the base feature set, in contrast, produced a significantly lower MAPE, at 53.54% (using a paired *t*-test, *p*-value = 0.003). These results are consistent with the daily prediction results.

Study 2: A Randomized Experiment on Causality and Its Mechanism

Motivation

The predictive results further motivated us to study the effects of eight types of discrete emotions in an experimental study, as Study 1 suggests that discrete emotions are more fundamental than valence and complex emotions in terms of understanding emotions in online reviews. An important question is whether the relationship between discrete emotions in online reviews and sales is causal or driven by confounding factors. If the relationship is causal, we can expect the predictive relationship to be robust. Otherwise, the predictive power may vanish when the confounding factors change. In this sense, a causal examination can complement the results of predictions (Hofman et al., 2017). More importantly, the predictive results of machine learning algorithms are not interpretable. An interpretable causal mechanism would complement our predictive study in several ways. First, from a theoretical perspective, it adds to the understanding of how discrete emotions in online reviews affect individuals' purchase intention. Second, understanding the mechanism would improve the external validity of our results. If the effects of emotions originate from enhancing review perceived processing fluency, which is not context-specific, the effects of emotions can be generalized to similar contexts, such as online reviews of restaurants, books, and music. Further, if a mediating effect exists, we can expect that the predictive relationship is stronger when consumers face information overload, and that the predictive relationship may be weaker when the information that consumers need to process is relatively simple. This also would generate additional managerial implications. For example, firms may decide whether to adopt predictive analytics with emotion analysis based on the level of information complexity in reviews that their targeted customers encounter. Another implication is that online review platforms may design features to make information processing more fluent, which contributes to review helpfulness. Finally, the effects of emotions may be different across cultures. Examining this relationship in the Chinese and U.S. movie markets can increase the cross-culture generalizability of our results.

[25] This result could be due to the following. First, although the Word2Vec representations may comprehensively contain semantic information not limited to emotional expressions, much of this semantic information may be irrelevant to consumer decisions, which merely adds noise to the prediction model. Second, the prediction task becomes challenging when directly adding word vectors to the model due to their high-dimensional nature. We had 499 movie samples, which is comparable to the dimensions of word vectors. Thus, the model performance may have been undermined due to dimensionality and overfitting. Third, when we constructed discrete emotion variables, we incorporated complementary information from the basic lexicon (Ren-CECps). The basic lexicon contains human-annotated information about how the word is associated with discrete emotions. Such information is not included in a word vector. This additional information has been discussed in cognitive psychological theories (Klauer, 1997; Schwarz, 1990; Van Kleef, 2009), and our experimental study shows that it has significant impacts on consumer decisions, thereby helping to improve prediction accuracy.

Stimulus Materials

To manipulate eight discrete emotions in online reviews, we prepared stimulus materials using real-world online movie reviews. As shown in Table 8, first, we identified two reviews without specific expression of the eight discrete emotions as positive or negative baseline reviews (Appendix D). Second, following the practice of Yin et al. (2014) and the definition of eight discrete emotions (Appendix E), we added emotional stimuli sentences at the beginning of the baseline reviews (the sentences in bold in Table 8) to manipulate the eight discrete emotions.

Procedure

We conducted a between-subject experiment, for which a total of 700 participants were hired from MTurk,[26] an online experiment platform. The 10 reviews in Table 8 define the 10 experimental groups, and each participant was randomly assigned to one group. The two groups with baseline reviews are the control groups, and the rest are the treatment groups. After removing the samples with an unreasonably short completion time (less than 4 seconds per question) and those that could not pass the attention check,[27] we achieved a valid sample of 634 participants. Of our participants, 59.9% were female; 53.1% were between 21 and 39 years old; 46.7% were 40 years old or older; 75.2% reported considering online reviews before they go to the movies, and 21.3% reported possibly considering online reviews before they go to the movies (we provide balance checks in Appendix F). After reading the review,[28] the participant rated (1) emotion embedded in reviews (Yin et al., 2014); (2) processing fluency (Storme et al., 2015); (3) perceived helpfulness of the reviews (Sen & Lerman, 2007); (4) purchase intention for the movie (Lee et al., 2014); and (5) perceived writer rationality (Xiao et al., 2018).[29] All questions were the same as in previous studies (Appendix G) and were measured using a 9-point Likert-type scale.

Results

We first conducted a manipulation check to ensure that each type of emotional content was successfully targeted, as shown in Table 9. In the two control groups, we expected no difference among the emotions (the null hypothesis). We conducted ANOVA tests on the two control groups (for positive baseline, $p = 0.342$, $n = 73$; for negative baseline, $p = 0.877$, $n = 51$), and the results show that we could not reject the null hypothesis. Further, in the eight treatment groups, the *t*-test results show that each treated emotion in the treatment group was significantly higher than that in the corresponding control group (ps < 0.10), indicating that the stimulus materials were effective. We present the reliability and validity check of three major constructs in Table 10. Cronbach's alphas for perceived processing fluency (PF), perceived review helpfulness (PH), purchase intention (PI), and perceived writer rationality (PR) are 0.92, 0.96, 0.97, and 0.81, respectively, indicating adequate construct reliability. Given that the four scales were used in previously validated research, we conducted confirmatory factor analysis to assess convergent and discriminant validity. The average variances extracted (AVEs) for the four constructs are all above 0.7, demonstrating convergent validity; the AVE of each of the constructs is higher than the highest squared correlation with any other latent variable (0.64), demonstrating discriminant validity (Fornell & Larcker, 1981).

Table 11 presents the results of the ordinary least square (OLS) estimation. To obtain robust standard errors, we allowed clustered standard errors at the group level. Each emotion variable in the table is a dummy variable that indicates whether a participant belonged to the corresponding treatment group and estimates the effect of the corresponding treatment condition. We controlled for participants' frequency of going to the cinema (*Freq*), age (*Age*), gender (*Gender*), and whether participants considered online reviews before purchasing a movie ticket (*Reviews*) in all regressions.

[26] We recruited only MTurk Master Workers in our study to ensure the quality of the responses. According to MTurk, "A Master Worker is a top Worker of the MTurk marketplace that has been granted the Mechanical Turk Masters Qualification. These Workers have consistently demonstrated a high degree of success in performing a wide range of human intelligence tasks across a large number of Requesters."

[27] We added an instruction for the attention check, "Please leave this question blank," to ensure the quality of the responses (Desimone et al., 2015).

[28] Consistent with the practice in IS research on online reviews, we asked the participants to read one review and then report outcomes (Xiao et al., 2018; Yin et al., 2014, 2017) because inviting the participants to read multiple reviews could induce confounding factors. For example, the recall of emotions in the last review may affect the effectiveness of emotions in the current review. Future research could investigate the case in which participants read a portfolio of emotional reviews.

[29] Xiao et al. (2018) studied perceived writer rationality and showed that it is relevant to anger in reviews. As we detail later, this helped us understand the mechanism of the effect of anger.

Table 8. Stimulus Materials

Group	Review
Positive baseline	The three-act narrative structure is clear and the second act is interesting, full of sly humor. The actor and actress's performance is good as well. The film succeeds in terms of character development.
Surprise	**I am very surprised after watching this movie.** The three-act narrative structure is clear and the second act is interesting, full of sly humor. The actor and actress's performance is good as well. The film succeeds in terms of character development.
Anticipation	**I anticipated the movie to be good before I watched it, and indeed,** the three-act narrative structure is clear and the second act is interesting, full of sly humor. The actor and actress's performance is good as well. The film succeeds in terms of character development.
Love	**I love the movie very much!** The three-act narrative structure is clear and the second act is interesting, full of sly humor. The actor and actress's performance is good as well. The film succeeds in terms of character development.
Joy	**It was very joyful to watch the movie!** The three-act narrative structure is clear and the second act is interesting, full of sly humor. The actor and actress's performance is good as well. The film succeeds in terms of character development.
Negative baseline	What kept this film from being great was the failure to show us the purpose behind the entire drive and desire to start this project in the first place. They didn't show what the end of the project had looked like, what the investors felt, how actress and her company ended up, and most of all how the two main characters continued with life.
Anger	**I got very angry after watching the movie!** What kept this film from being great was the failure to show us the purpose behind the entire drive and desire to start this project in the first place. They didn't show what the end of the project had looked like, what the investors felt, how actress and her company ended up, and most of all how the two main characters continued with life.
Anxiety	**I felt very anxious toward the movie. If the director makes another movie like this, I'd be concerned!** What kept this film from being great was the failure to show us the purpose behind the entire drive and desire to start this project in the first place. They didn't show what the end of the project had looked like, what the investors felt, how actress and her company ended up, and most of all how the two main characters continued with life.
Sadness	**I felt very sad after watching the movie.** What kept this film from being great was the failure to show us the purpose behind the entire drive and desire to start this project in the first place. They didn't show what the end of the project had looked like, what the investors felt, how actress and her company ended up, and most of all how the two main characters continued with life.
Disgust	**I felt very disgusted after watching the movie!** What kept this film from being great was the failure to show us the purpose behind the entire drive and desire to start this project in the first place. They didn't show what the end of the project had looked like, what the investors felt, how actress and her company ended up, and most of all how the two main characters continued with life.

Table 9. Manipulation Check

Emotion	Mean of control group	Mean of treatment group	*p*-value of *t*-test
Surprise	5.685 (73)	6.362 (69)	0.052 (142)
Anticipation	6.123 (73)	6.985 (68)	0.004 (141)
Love	5.877 (73)	6.576 (59)	0.035 (132)
Joy	6.274 (73)	6.767 (60)	0.073 (133)
***p*-value of ANOVA**	**0.342 (73)**	**--**	**--**
Anger	5.510 (51)	7.629 (62)	0.000 (113)
Anxiety	5.451 (51)	6.576 (66)	0.000 (117)
Sadness	5.490 (51)	6.918 (61)	0.001 (112)
Disgust	5.804 (51)	7.292 (65)	0.001 (116)
***p*-value of ANOVA**	**0.877 (51)**	**--**	**--**

Note: Sample size in parentheses.

Table 10. Reliability and Validity of Constructs

Construct	Cronbach's alpha	AVE in CFA	Squared correlation of factors			
			PF	PH	PI	PR
PF	0.92	0.78	1.00			
PH	0.95	0.88	0.52	1.00		
PI	0.97	0.91	0.32	0.42	1.00	
PR	0.81	0.77	0.50	0.64	0.26	1.00

Note: PF = Perceived processing fluency; PH = Perceived review helpfulness; PI = Purchase intention; AVE = Average variance extracted; PR = Perceived writer rationality; CFA = Confirmatory factor analysis.

Table 11. Regression Results

Variable	(1) DV: PI	(2) DV: PH	(3) DV: PF	(4) DV: PH	(5) DV: PI	(6) DV: PH	(7) DV: PF	(8) DV: PH	(9) DV: PR	(10) DV: PH	(11) DV: PI
Anticipation	0.48*** (0.11)	0.40** (0.16)	1.02*** (0.08)	-0.02 (0.13)							
Joy	0.24*** (0.08)	0.13 (0.09)	0.78*** (0.08)	-0.19** (0.08)							
Love	1.22*** (0.11)	0.94*** (0.11)	1.34*** (0.09)	0.38*** (0.10)							
Surprise	0.86*** (0.12)	0.93*** (0.08)	0.93*** (0.10)	0.54*** (0.08)							
Anger					-2.49*** (0.30)	-0.53** (0.24)	0.34* (0.18)	-0.67*** (0.19)	-0.49*** (0.05)	-0.61 (0.37)	
Anxiety					-0.53*** (0.02)	-0.63*** (0.04)	-0.46*** (0.02)	-0.44*** (0.04)			
Disgust					-2.48*** (0.26)	-1.67*** (0.22)	-0.54*** (0.18)	-1.44*** (0.16)			
Sadness					-2.60*** (0.28)	-0.39* (0.20)	-0.28* (0.16)	-0.27* (0.14)			
PH											0.41*** (0.15)
PF				0.42*** (0.07)				0.42*** (0.04)			
PR										0.53** (0.24)	
Frequency	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reviews	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	5.08*** (0.19)	5.38*** (0.14)	6.47*** (0.16)	2.68*** (0.51)	5.80*** (0.63)	5.63*** (0.60)	5.29*** (0.29)	3.40*** (0.48)	5.92*** (0.13)	4.44*** (1.28)	2.69*** (0.58)
N	329	329	329	329	305	305	305	305	113	113	634
Adj. R^2	0.03	0.05	0.05	0.21	0.33	0.12	0.06	0.31	0.11	0.16	0.16

Note: Each emotion variable is a dummy variable that indicates whether a participant belongs to the corresponding treatment group; PI = Purchase intention, PH = Perceived review helpfulness, PF = Perceived processing fluency, PR = Perceived writer rationality. Frequency, Age, Gender, and Reviews control for subjects' frequency of movie watching, ages, genders, and whether they consider online reviews before purchasing a movie ticket, respectively. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

In Columns 1 to 4 (Table 11), we used the samples in the positive baseline group and four positive emotion groups (n = 329). Columns 1 to 3 show that each positive emotion has a positive effect on PI, PH, and PF, respectively. Other than joy on PH, all coefficients are statistically significant. Online reviews with joyful expressions may be perceived as less helpful because overt expressions of joy in content tend to signal the writer's overconfidence (Jiang et al., 2019). Column 4 shows the result of adding the mediator PF to the regression of PH. Compared to the results in Column 2, all coefficients of positive emotions in Column 4 are smaller, indicating that PF mediates the effects of all positive emotions on PH (Baron & Kenny, 1986). In Columns 5 to 8, we use the samples in the negative baseline group and four negative emotion groups (n = 305). Columns 5 to 7 show that each negative emotion has a significantly negative effect on PI, PH, and PF (except anger on PF), respectively. Column 8 shows the result of adding the mediator PF to the regression of PH. Compared to Column 6, all coefficients of negative emotions in Column 8 are smaller in effective size (except anger), indicating that PF mediates the effects of anxiety, sadness, and disgust on PH. Anger is different from other negative emotions because it reflects the individual's irrationality and low coping efforts (Kalamas et al., 2008; Xiao et al., 2018). Review readers are likely to

attribute anger to the writer's irrationality and then stop cognitively processing the review information (Kim & Gupta, 2012; Xiao et al., 2018). This is consistent with our experimental results, in which anger increased perceived processing fluency but decreased perceived review helpfulness, as the reader may quickly judge the angry review as of low informative value. To further confirm this deduction, we conducted an additional analysis with samples in the anger group and negative baseline group. Column 9 shows that indeed, anger negatively affects PR. Column 10 shows that when adding the mediator PR to the regression of PH on anger, the effect of anger becomes insignificant, indicating a complete mediation effect of PR on anger (Baron and Kenny 1986). Column 11 confirms the positive effect of PH on PI. Notably, because our analysis contains multiple statistical inferences simultaneously, the risk of a Type 1 error (false significance) increases. To address this concern, we conducted a Bonferroni correction (Simes, 1986). The results show that our findings are robust (detailed in Appendix H). Finally, we normalized the treatment effect of each discrete emotion on PI and PH to the same scale, as shown in Table 1. The normalized effect of surprise, anticipation, love, joy, anger, anxiety, sadness, and disgust on PI (PH) is 1.27 (1.38), 0.55 (0.47), 1.75 (1.35), 0.49 (0.26), -1.18 (-0.25), -0.47 (-0.56), -1.82 (-0.28), and -1.66 (-1.12), respectively. All effect sizes are higher than the existing nonemotional factors listed in Table 1, which demonstrates the economic importance of discrete emotions in online reviews.

We review the literature to derive theoretical insight into the heterogeneous effects of emotions on PI. First, because movie consumers tend to be arousal seeking (Xie & Lee, 2008), high-arousal positive emotional experiences (i.e., love and surprise) associated with movie consumption are more rewarding than low-arousal ones (i.e., joy and anticipation). Similarly, high-arousal negative emotional experiences (i.e., anxiety and anger) are less detrimental to PI than low-arousal ones (i.e., sadness and disgust). Second, low-arousal positive/negative emotions are of similar effectiveness, but high-arousal positive/negative emotions are not. The difference between high-arousal emotions can be explained based on extant research. Love is an emotion implying high levels of certainty, trust, and acceptance (Lazarus, 1991; Plutchik, 1984), whereas surprise is a more neutral emotion compared to love (Nguyen et al., 2020). Therefore, writers' expressions of love tend to generate a greater endorsement of a movie than expressions of surprise. Expressions of anger are especially likely to be noticed, encoded, recalled, and can directly affect consumers' attitudes (Yin et al., 2021). Anxiety in reviews tends to increase the perceived uncertainty of product quality (Yin et al., 2014), which may only indirectly affect purchase intention through readers' risk-averse tendencies. Therefore, anger tends to be more effective in decreasing purchase intention than anxiety.

Discussion

Our work makes several contributions to the information systems literature in regard to emotions in online reviews and sales. First, we find that discrete emotions, as compared to valence (Hennig-Thurau et al., 2015; Song et al., 2019) and latent emotion topics (Yu et al., 2012), better represent emotional information embedded in online reviews for predicting sales. We also demonstrate that discrete emotions are substitutive for valence and latent emotion topics. With discrete emotion analysis outperforming valence analysis, the prevalent practice in industry and academia, our work suggests that researchers and managers can use discrete emotions to advance the understanding of consumer emotions embedded in large-scale online user-generated content. Our results further provide empirical evidence that supports discrete emotion theory (Lerner et al., 2015; Plutchik & Kellerman, 1982; Tomkins, 1962). Our finding that theory-driven discrete emotional features are more informative than data-driven latent emotion topics is also valuable.

Second, our work demonstrates the economic importance of discrete emotions in online reviews. We show that the effectiveness of discrete emotions is comparable to or higher than that of the existing factors in the literature, including valence, arousal (East et al., 2017; Kim & Gupta, 2012; Yin et al., 2017), review quality, credibility (Teng et al., 2017), and reviewer characteristics (Huang et al., 2015; Ngo-Ye & Sinha, 2014; Yin et al., 2016). Further, we are among the first to show that the effectiveness of discrete emotions in online reviews is robust across different cultures.

Third, our work takes an initial but important step toward understanding the effects of discrete emotions from a new perspective that has been overlooked in the information systems literature, i.e., the mediating effect of perceived processing fluency. Differing from the existing mechanisms, i.e., perceived writer cognitive effort (Yin et al., 2014) and perceived writer rationality (Xiao et al., 2018), which explain only the effects of anxiety and anger, the mediating effect of processing fluency is prevalent in all types of discrete emotions, except anger. Future research might involve more comprehensive investigations of this theoretically interesting and practically important topic. For example, it would be valuable to determine how to use empirical methods to predict perceived processing fluency in reviews and to determine whether processing fluency is associated with other factors beyond emotions in online reviews, e.g., the writer's review writing ability.

Fourth, through advancing existing work (Xue et al., 2014), our work proposes an approach that enables automatic domain-adaptive emotion lexicon construction and multidimensional emotion detection in texts, which adds to the literatures on affective computing (Picard, 2003) and domain-driven data mining (Cao, 2010).[30] We construct and validate a new emotion lexicon specific to microblogging movie reviews for future research and application.[31]

This work has several managerial implications. First, this paper highlights the economic importance of discrete emotions in online reviews. Online review platforms, such as Amazon, Yelp, and IMDB, currently ask users to rate only their positive/negative attitudes (i.e., valence) toward consumption experiences. Our work calls for attention to design features that would facilitate discrete emotional expressions.[32]

Second, the mediating effect of processing fluency can be leveraged to enable writers to write helpful reviews. When a platform detects that a consumer is writing a positive review with low readability (e.g., of excessive length or with information presented in a complicated way), it could suggest that the consumer should more explicitly express their positive emotions to increase review processing fluency. When a platform detects that a consumer is writing a negative review, it could advise the consumer to write a highly readable review (with more rational expression if the consumer's emotion is anger), instead of using more negative and emotional expression. Further, platforms could leverage our algorithm to add emotional tags to highlight positive discrete emotions in reviews, which could enhance the readers' fluency in processing review information.

Third, our results can help movie marketers develop online movie marketing strategies. Our methodology enables marketers to detect online reviews with different discrete emotions. Thus, to boost box office sales, movie marketers could shift resources to design, share, reply to, or control online reviews in regard to their embedded emotions (Song et al., 2019). In addition, because our methodology involves only public data, it could help marketers predict the performance of other competitive movie products that are released in the same period and facilitate competitive responses.

Fourth, our approach can help cinema managers predict box office sales more accurately, thereby helping them to optimally select movies, arrange screenings (Song et al., 2019; Yu et al., 2012), and make operational decisions on inventory management. Finally, our methodology could be easily adapted to other domains (e.g., books, music, restaurants, car sales) to detect emotions in consumer reviews and predict future business performance.

We acknowledge some limitations of this work, which could motivate future research. First, we did not assess the recall of the emotion words in our lexicon because it is extremely difficult to obtain the ground truth of how many unique emotion words there are in 3.26 million microblogging messages. Second, in Study 2, we used participants' self-reported data to measure purchase intention. To mitigate self-report biases, future research could offer participants the option of buying a movie ticket as a means of observing their actual purchase behaviors. Third, as online platforms, including Facebook, are adopting more predesigned emotion-expression options for users, researchers could develop an experiment to evaluate the business value of such emotion-expression-related designs. Fourth, the mechanisms through which perceived processing fluency affects perceived review helpfulness could be further explored. We found evidence indicating that processing fluency is positively associated with perceived writer rationality, effort, and review informativeness (Appendix I). Finally, we used a Word2Vec model to map words into vectors. Future research might investigate whether more sophisticated models, e.g., transformer-based models, could produce more accurate word vectors and thereby generate better emotion lexicons. Further, in Algorithm 2, we used a weighted average of the emotional intensities of the most similar words to predict the emotional intensities of the newly extended words. The weights are based on semantic similarity, which is reasonable but not necessarily optimal. The attention-based architecture could be leveraged to learn the optimal weights.

[30] In particular, the previous work by Xue et al. (2014) can be viewed as a special case of our Algorithm 1 when f is implemented by a K-nearest neighbor algorithm, without the iterative process, and when there are only two emotional dimensions (positive and negative). We contribute to their work as follows. First, we modify f by introducing another hyperparameter α to achieve better accuracy. Further, we show that using a KNN-type algorithm that uses the cosine similarity between word vectors achieves better performance than directly training the state-of-the-art models on the 200-dimension word vectors (as we detail in Appendix C). Second, we propose an iterative process that enables us to mine as many words as possible in the domain lexicon. Our method detected 30.2% (1,556 out of 5,154) more emotion words than theirs. More importantly, we show that test errors will remain consistently low as the number of iterations grows, which demonstrates the validity of the process. Third, we validate the newly mined words with out-of-sample testing and human annotation. Fourth, we propose a more general framework for lexicon extension with statistical language modeling.

[31] The lexicon is available at https://drive.google.com/file/d/1s5vXmiIlby0qZldnqzmjHSDKtcKpL39j/view?usp=sharing.

[32] Two notable exceptions that consider such features are (1) Facebook.com, which had a major expansion of its emoji responses from "Like" to multiple emotions/emojis (e.g., angry, funny, love, like), and (2) Rappler.com, a news website that at the end of each piece of news, asked its audience to express their feelings by choosing one out of eight discrete emotions.

Acknowledgments

The authors thank the senior editor, the associate editor, and the anonymous reviewers for constructive suggestions. Jinghua Huang is the corresponding author. The earlier versions of this work were presented at the 14th INFORMS Workshop on Data Mining and Decision Analytics (DMDA) 2019, China Summer Workshop on Information Management (CSWIM) 2019, China Marketing International Conference (CMIC) 2019, and 16th INFORMS Annual Meeting (INFORMS) 2021. This research was supported by the National Natural Science Foundation of China [Grant:72072100].

References

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the International Conference on Learning Representations*.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the International Conference on Web Intelligence* (pp. 492-499). https://doi.org/10.1109/WI-IAT.2010.63

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research. conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182. https://doi.org/10.1037/0022-3514.51.6.1173

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: controversies and emerging consensus. *Current Directions in Psychological Science*, *8*(1), 10-14. https://doi.org/10.1111/1467-8721.00003

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(1), 281-305. https://dl.acm.org/doi/10.5555/2188385.2188395

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022. https://dl.acm.org/doi/10.5555/944919.944937

Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. In A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, & V. C. Müller (Eds.), *Cognitive Behavioural Systems* (pp. 144-157). Springer. https://doi.org/10.1007/978-3-642-34584-5_11

Cao, L. (2010). Domain-driven data mining: Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 755-769. https://doi.org/10.1109/TKDE.2010.32

Chen, R., & Sakamoto, Y. (2016). The effect of fluency on review helpfulness: Does it depend on perspective-taking? In *Proceedings of the Pacific Asia Conference on Information Systems*. https://aisel.aisnet.org/pacis2016/385

Chen, R., & Zhang, J. (2018). The effects of fluency and framing on perceived review helpfulness. In *Proceedings of 2018 Global Marketing Conference* (pp. 732-736).

Desimone, J. A., Harms, P. D., & Desimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171-181. https://doi.org/10.1002/job.1962

Desmet, P. (2010). Three levels of product emotion. In *Proceedings of Kansei Engineering and Emotion Research International Conference*.

DeSteno, D., Petty, R. E., Wegener, D. T., & Rucker, D. D. (2000). Beyond valence in the perception of likelihood: The role of emotion specificity. *Journal of Personality and Social Psychology*, *78*(3), 397-416. https://doi.org/10.1037/0022-3514.78.3.397

East, R., Romaniuk, J., Chawdhary, R., & Uncles, M. (2017). The impact of word of mouth on intention to purchase currently used and other brands. *International Journal of Market Research*, *59*(3), 321-334. https://doi.org/10.2501/IJMR-2017-026

Egeth, H., & Kahneman, D. (1975). Attention and effort. *The American Journal of Psychology*, *88*(2), 339. https://doi.org/10.2307/1421603

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology*, *81*(5), 869-885. https://doi.org/10.1037/0022-3514.81.5.869

Felbermayr, A., & Nanopoulos, A. (2016). The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing*, *36*, 60-76. https://doi.org/10.1016/j.intmar.2016.05.004

Filieri, R., McLeay, F., Tsui, B., & Lin, Z. (2018). Consumer perceptions of information helpfulness and determinants of purchase intention in online consumer reviews of services. *Information and Management*, *55*(8), 956-970. https://doi.org/10.1016/j.im.2018.04.010

Fornell, C., & Larcker, D. F. (1981). Evaluating structural model with unobserved variables and measurement errors. *Journal of Marketing Research*, *18*(1), 39-50. https://doi.org/10.2307/3151312

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forums and search for sales prediction of high-involvement products. *MIS Quarterly*, *41*(1), 65-82. https://doi.org/10.25300/MISQ/2017/41.1.04

Havlena, W. J., & Holbrook, M. B. (1986). The varieties of consumption experience: Comparing two typologies of emotion in consumer behavior. *Journal of Consumer Research*, *13*(3), 394-394. https://doi.org/10.1086/209078

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, *43*(3), 375-394. https://doi.org/10.1007/s11747-014-0388-3

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324). https://doi.org/10.1126/science.aal3856

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70. https://www.jstor.org/stable/4615733

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, *48*, 17-27. https://doi.org/10.1016/j.chb.2015.01.010

Huang, L., Tan, C. H., Ke, W., & Wei, K. K. (2014). Do we order product review information display? How? *Information and Management*, *51*(7), 883-894. https://doi.org/10.1016/j.im.2014.

05.002

Isbell, L. M., Lair, E. C., & Rovenpor, D. R. (2013). Affect-as-information about processing styles: A cognitive malleability approach. *Social and Personality Psychology Compass*, *7*(2), 93-114. https://doi.org/10.1111/spc3.12010

Jiang, L., Yin, D., & Liu, D. (2019). Can joy buy you money? The impact of the strength, duration, and phases of an entrepreneur's peak displayed joy on funding performance. *Academy of Management Journal*, *62*(6), 1848-1871. https://doi.org/10.5465/amj.2017.1423

Kalamas, M., Laroche, M., & Makdessian, L. (2008). Reaching the boiling point: Consumers' negative affective reactions to firm-attributed service failures. *Journal of Business Research*, *61*(8), 813-824. https://doi.org/10.1016/j.jbusres.2007.09.008

Kenter, T., Borisov, A., & De Rijke, M. (2016). Siamese CBOW: Optimizing word embeddings for sentence representations. In *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/ 10.18653/v1/p16-1089

Kim, J., & Gupta, P. (2012). Emotional expressions in online user reviews: How they influence consumers' product evaluations. *Journal of Business Research*, *65*(7), 985-992. https://doi.org/10.1016/j.jbusres.2011.04.013

Klauer, K. C. (1997). Affective priming. *European Review of Social Psychology*, *8*(1), 67-103. https://psycnet.apa.org/record/1998-06032-003

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, *42*(7), 3751-3759. https://doi.org/10.1016/j.eswa.2014.12.044

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, *46*(4), 352-352. https://doi.org/10.1037/0003-066X.46.4.352

Lee, K., Choi, J., & Li, Y. J. (2014). Regulatory focus as a predictor of attitudes toward partitioned and combined pricing. *Journal of Consumer Psychology*, *24*(3), 355-362. https://doi.org/10.1016/j.jcps.2014.01.001

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*(1), 799-823. https://doi.org/10.1146/annurev-psych-010213-115043

Lerner, J. S., Small, D. A., & Loewenstein, G. (2004). Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science*, *15*(5), 337-341. https://doi.org/10.1111/j.0956-7976.2004.00679.x

Li, M., Huang, L., Tan, C. H., & Wei, K. K. (2013). Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, *17*(4), 101-136. https://doi.org/10.2753/JEC1086-4415170404

Liu, A. X., Xie, Y., & Zhang, J. (2019). It's not just what you say, but how you say it: The effect of language style matching on perceived quality of consumer reviews. *Journal of Interactive Marketing*, *46*(1), 70-86. https://doi.org/10.1016/j.intmar.2018.11.001

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74-89. https://doi.org/10.1509/jmkg.70.3.74

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review*, *90*(2), 426-432. https://doi.org/10.5815/ijigsp.2012.01.06

Malik, M. S. I., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, *73*(1), 290-302. https://doi.org/10.1016/j.chb.2017.03.053

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185-200. https://dl.acm.org/doi/10.5555/2017447.2017457

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, *61*(1), 47-58. https://doi.org/10.1016/j.dss.2014.01.011

Nguyen, H., Calantone, R., & Krishnan, R. (2020). Influence of social media emotional word of mouth on institutional investors' decisions and firm value. *Management Science*, *66*(2), 887-910. https://doi.org/10.1287/mnsc.2018.3226

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*(1), 62-73. https://doi.org/10.1016/j.dss.2016.02.013

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 115-124). https://doi.org/10.3115/1219840.1219855

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies*, *59*(1), 55-64. https://doi.org/10.1016/S1071-5819(03)00052-1

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 197-219). Psychology Press.

Plutchik, R. (2001). The nature of emotions. *American Scientist*, *89*(4), 344-350. https://www.jstor.org/stable/27857503

Plutchik, R., & Kellerman, H. (1982). Emotions are more than subjective experiences. *Contemporary Psychology*, *27*(7), 581-581. https://doi.org/10.1037/021362

Quan, C., & Ren, F. (2010). A blog emotion corpus for emotional expression analysis in Chinese. *Computer Speech and Language*, *24*(4), 726-749. https://doi.org/10.1016/j.csl.2010.02.002

Reber, R., Fazendeiro, T. A., & Winkielman, P. (2002). Processing fluency as the source of experiences at the fringe of consciousness. *Psyche: An Interdisciplinary Journal of Research on Consciousness*, *8*(10), 1-21. https://psycnet.apa.org/record/2002-04734-001

Risselada, H., de Vries, L., & Verstappen, M. (2018). The impact of social influence on the perceived helpfulness of online consumer reviews. *European Journal of Marketing*, *52*(3), 619-636. https://doi.org/10.1108/EJM-09-2016-0522

Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? The effect of tweets on movie sales. *Decision Support Systems*, *55*(4), 863-870. https://doi.org/10.1016/j.dss.2012.12.022

Russell, J. A., & Mehrabian, A. (1974). Distinguishing anger and anxiety in terms of emotional response factors. *Journal of Consulting and Clinical Psychology*, *42*(1), 79-83. https://doi.org/10.1037/h0035915

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., Sams, M., & Nummenmaa, L. (2016).

Discrete neural signatures of basic emotions. *Cerebral Cortex*, *26*(6), 2563-2573. https://doi.org/10.1093/cercor/bhv086

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer et al. (Eds.), *Appraisal processes in emotion: Theory, Methods, Research* (pp. 92-120). Oxford University Press.

Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: the role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234-243. https://doi.org/10.1002/cb.1372

Schwarz, N. (1990). *Feelings as information: Informational and motivational functions of affective states*. The Guilford Press.

Sen, S., & Lerman, D. (2007). Why are you telling me this? An examination into negative consumer reviews on the web. *Journal of Interactive Marketing*, *21*(4), 76-94. https://doi.org/10.1002/dir.20090

Shi, W. (2017). *Big data mining of Chinese Weibo texts: Sentiment analysis perspective*. China Social Sciences Press.

Shimp, T. A., & Stuart, E. W. (2004). The role of disgust as an emotional mediator of advertising effects. *Journal of Advertising*, *33*(1), 43-53. https://doi.org/10.1080/00913367.2004.10639150

Shiv, B. (2007). Emotions, decisions, and the brain. *Journal of Consumer Psychology*, *17*(3), 174-178. https://doi.org/10.1016/S1057-7408(07)70025-6

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751-754. https://doi.org/10.1093/biomet/73.3.751

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813-838. https://doi.org/10.1037/0022-3514.48.4.813

So, J., Achar, C., Han, D. H., Agrawal, N., Duhachek, A., & Maheswaran, D. (2015). The psychology of appraisal: Specific emotions and decision-making. *Journal of Consumer Psychology*, *25*(3), 359-371. https://doi.org/10.1016/j.jcps.2015.04.003

Song, T., Huang, J., Tan, Y., & Yu, Y. (2019). Using user- and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Information Systems Research*, *30*(1), 191-203. https://doi.org/10.1287/isre.2018.0797

Song, Y., Shi, S., Li, J., & Zhang, H. (2018). Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/N18-2028

Storme, M., Myszkowski, N., Davila, A., & Bournois, F. (2015). How subjective processing fluency predicts attitudes toward visual advertisements and purchase intention. *Journal of Consumer Marketing*, *32*(6), 432-440. https://doi.org/10.1108/JCM-10-2014-1187

Teng, S., Khong, K. W., Chong, A. Y. L., & Lin, B. (2017). Examining the impacts of electronic word-of-mouth message on consumers' attitude. *Journal of Computer Information Systems*, *57*(3), 238-251. https://doi.org/10.1080/08874417.2016.1184012

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer.

Tsekouras, D. (2017). The effect of rating scale design on extreme response tendency in consumer product ratings. *International Journal of Electronic Commerce*, *21*(2), 270-296. https://doi.org/10.1080/10864415.2016.1234290

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science*, *18*(3), 184-188. https://doi.org/10.1111/j.1467-8721.2009.01633.x

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, *22*(12), 2864-2885. https://doi.org/10.1162/jocn.2009.21366

Watson, L., & Spence, M. T. (2007). Causes and consequences of emotions on consumer behaviour: A review and integrative cognitive appraisal theory. *European Journal of Marketing*, *41*(5-6), 487-511. https://doi.org/10.1108/03090560710737570

Xiao, Y., Zhang, H., & Cervone, D. (2018). Social functions of anger: A competitive mediation model of new product reviews. *Journal of Product Innovation Management*, *35*(3), 367-388. https://doi.org/10.1111/jpim.12425

Xie, G. X., & Lee, M. J. (2008). Anticipated violence, arousal, and enjoyment of movies: Viewers' reactions to violent previews based on arousal-seeking tendency. *The Journal of Social Psychology*, *148*(3), 277-292. https://doi.org/10.3200/SOCP.148.3.277-292

Xue, B., Fu, C., & Shaobin, Z. (2014). A study on sentiment computing and classification of Sina Weibo with word2vec. In *Proceedings of the IEEE International Congress on Big Data*, (pp. 358-363). https://doi.org/10.1109/BigData.Congress.2014.59

Yang, X., Zhang, Z., Zhang, Z., Mo, Y., Li, L., Yu, L., & Zhu, P. (2016). Automatic construction and global optimization of a multisentiment lexicon. *Computational Intelligence and Neuroscience*, Article 2093406. https://doi.org/10.1155/2016/2093406

Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539-560. https://doi.org/10.25300/MISQ/2014/38.2.10

Yin, D., Bond, S. D., & Zhang, H. (2017). Keep your cool or let it out: Nonlinear effects of expressed arousal on perceptions of consumer reviews. *Journal of Marketing Research*, *54*(3), 447-463. https://doi.org/10.1509/jmr.13.0379

Yin, D., Bond, S. D., & Zhang, H. (2021). Anger in consumer reviews: Unhelpful but persuasive? *MIS Quarterly*, *45*(3), 1059-1086. https://doi.org/10.25300/MISQ/2021/15363

Yin, D., Mitra, S., & Zhang, H. (2016). When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. *Information Systems Research*, *27*(1), 131-144. https://doi.org/10.1287/isre.2015.0617

Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720-734. https://doi.org/10.1109/TKDE.2010.269

Zeelenberg, R., Wagenmakers, E. J., & Rotteveel, M. (2006). The impact of emotion on perception: Bias or enhanced processing? *Psychological Science*, *17*(4), 287-291. https://doi.org/10.1111/j.1467-9280.2006.01700.x

About the Authors

Yifan Yu is a Ph.D. candidate in the Department of Information Systems and Operations Management, University of Washington. He is an incoming assistant professor in the Department of Information, Risk, and Operations Management, The University of Texas at Austin. His research interests include emotion artificial intelligence, deep learning for unstructured data analytics, and the economics of social media networks. He has published in *Information Systems Research*. ORCiD: https://orcid.org/0000-0002-8959-3169.

Yang Yang received her Ph.D. from the Department of Business Management, School of Management, University of Science and Technology of China. She was a visiting scholar at the University of Science and Technology of China-UW (USTC-UW) Institute for Global Business and Finance Innovation. Her research interests include online emotional content, AI chatbot, industry-university-research collaboration. ORCiD: https://orcid.org/0000-0001-8694-1167.

Jinghua Huang is a full professor in the Research Center for Contemporary Management, Department of Management Science and Engineering, School of Economics and Management, Tsinghua University. She received her Ph.D. from Tsinghua University. Her research interests include the business value of IS/IT, social networking, and electronic business. Professor Huang has published papers in the following journals: *Information Systems Research*, *Journal of the Association for Information Systems*, *Information & Management*, *Communications of Association for Information Systems*, and *Electronic Commerce Research and Applications*, among others. Professor Huang is the corresponding author of this research. ORCiD: https://orcid.org/0000-0003-4344-7884.

Yong Tan is the Michael G. Foster Endowed Professor in Information Systems at the Michael G. Foster School of Business, University of Washington, and a Distinguished Academic Fellow of the INFORMS Information Systems Society. His research interests include social media and networks, sharing economy, fintech, mobile and electronic commerce, and big data analytics. He has published in *Information Systems Research*, *Management Science*, and *Management Information Systems Quarterly*, among other outlets. ORCiD: https://orcid.org/0000-0001-8087-3423.

Appendix A

Training Word2Vec Model

We trained a Word2Vec model by using 3.26 million Chinese movie-related microblogging messages and 5 million microblogging messages from the NLPIR microblog corpus. First, the Chinese word segmentation module in Python segments all of the training text into words. Second, the Word2Vec model constructs a dictionary for all unique words. Every word is represented by a one-hot vector. The vector has the same dimensions as the dictionary, with only one nonzero element (equal to 1), representing the corresponding word. Third, every one-hot vector is input into a multilayer neural network, whose output is the probabilistic distribution of the words adjacent to the input word. The initial distribution is randomly assigned. The model learns from the corpus and optimizes its weights so that the predicted probabilistic distribution ultimately approximates the observed distribution. Finally, for every word in the dictionary, we can extract its corresponding weight vector in the embedding layer (i.e., the first layer of the neural network). The weight vector, called a *word vector*, contains the semantic information of the word (Mikolov et al., 2013).

Appendix B

Random Hyperparameter Search

In Table B1, we present the results of the random hyperparameter search detailed in Algorithm 3. We find that the optimal parameter is K = 5 and α = 0.15, and the corresponding validation MAE is 0.07155. We note that all validation MAEs are between 0.071 to 0.077, indicating that the performance of Algorithm f is not very sensitive to hyperparameters.

Table B1. Hyperparameters and Validation MAEs

K	α	Validation error	K	α	Validation error
5	**0.15**	**0.07155**	3	0.20	0.07274
5	0.06	0.07156	4	0.20	0.07293
4	0.12	0.07156	2	0.10	0.07325
3	0.10	0.07186	2	0.08	0.07328
5	0.18	0.07192	2	0.06	0.07337
8	0.16	0.07193	2	0.00	0.07345
3	0.08	0.07202	8	0.20	0.07355
3	0.18	0.07219	9	0.20	0.07376
4	0.02	0.07220	10	0.20	0.07397
9	0.16	0.07244	1	0.14	0.07679
7	0.18	0.07248	1	0.04	0.07694
5	0.20	0.07249	1	0.00	0.07696
7	0.00	0.07251			

Appendix C

Comparison of Alternative Implementations of Algorithm f [redacted]

Alternatively, our algorithm f in Algorithm 2 can be constructed by using the mean model and more sophisticated state-of-the-art models, i.e., RF and XGB. For the mean model, the prediction of a new word is produced by using the average emotion intensities of all words in the training set, i.e., $f_{mean}(w; L_0^{tr}) = \sum_i \frac{v_i}{|L_0^{tr}|}$, where $v_i \in L_0^{tr}$.

Algorithm 4. Mapping Word Vectors to Emotional Intensities and Testing	
Pseudocode	**Remarks**
Input a basic lexicon $L_0 = \{(w_i, v_i)\}_i^N$;	L_0 is the same as in Algorithm 1.
Randomly split L_0 into training (80%), validation (10%), and test (10%) sets: L_0^{tr}, L_0^{va}, and L_0^{te};	
Input a pre-trained Word2Vec model;	
FOR each emotion dimension k ($k \in \{1,2,\ldots,8\}$):	
Initiate a machine learning model g_k;	In our case, g_k can be RF or XGB. We use the default setting suggested by *sklearn*.
FOR each word $w_i \in L_0^{tr}$:	
$y_i \leftarrow$retrieve the true value of the kth emotional intensity e_k^0 from v_i;	
$x_i \leftarrow$retrieve the word vector of w_i;	x_i is a 200-dimension vector.
END FOR	
Train g_k with data $D_k = \{(x_i, y_i)\}_{i=1}^{\lvert L_0^{tr}\rvert}$;	
END FOR	
FOR each word $w_i \in L_0^{te}$:	Calculate test errors with the test set.
Retrieve the true value of the emotional intensities $v_i = (e_1^0, e_2^0, \ldots, e_8^0)$;	
$x_i \leftarrow$retrieve the word vector of w_i;	
FOR $k \in \{1,2,\ldots,8\}$:	
$e_k \leftarrow g_k(x_i)$;	
END FOR	
Calculate MAE $\epsilon_i = \sum_{i=1}^{8} \frac{\lvert e_i - e_i^0\rvert}{8}$;	
END FOR	
Calculate average MAE $\epsilon_g = \sum_i \frac{\epsilon_i}{\lvert L_0^{te}\rvert}$;	
RETURN ϵ_g;	

The mean model serves as a benchmark for other models. In addition, we note that Algorithm 2 is based on using the semantic similarity information embedded in the cosine distance between word vectors. Alternatively, one may directly utilize the 200-dimension word vectors as features to predict the emotional intensities of the word. As detailed in Algorithm 4, we use two state-of-the-art models, RF and XGB, to implement this idea. We find that the out-of-sample MAEs of the mean model, RF, XGB, and our model (Algorithm 2) are 0.117, 0.094, 0.091, and 0.073, respectively. The results show that although RF and XGB can outperform the mean model, our algorithm achieves the best performance.

Appendix D

Preparation of Baseline Reviews

To identify reviews that were positive or negative in valence, but without specific expressions of the eight discrete emotions, we focused on 11 movies (*Wonder Park*, *Five Feet Apart*, *No Manches Frida 2*, *Captive State*, *Dominirriquenos 2*, *The Mustang*, *Faith, Hope and Love*, *The Aftermath*, *Ash is Purest White*, *The Hummingbird Project*, and *Combat Obscura*) that were released on March 15, 2019. We scanned all 706 reviews of the 11 movies on IMDb and deleted redundant reviews and those of extreme length. We removed (1) the movie title, (2) emoji images, and (3) the emotional content that directly indicates reviewers' discrete emotions, such as "I love this movie," and "Disgusting experience," and collected 18 non-emotional reviews as baseline samples (nine positive baseline reviews and nine negative baseline reviews) (Yin et al., 2014). Then, 50 participants were hired from MTurk to rate the emotions in reviews, based on the following question, "In your opinion, to what extent does each of the following words describe how the reviewer felt when he/she wrote the above review? (Surprise, Anticipation, Love, Joy, Anger, Anxiety, Disgust, Sadness), using a 9-point Likert-type scale of 1 = not at all to 9 = very much) (Yin et al., 2014). The results of an ANOVA show that the eight reviews demonstrate no significant difference among the various discrete emotions ($ps > 0.052$). The authors discussed the eight reviews and chose two (one positive baseline review and one negative baseline review) that are feasible in regard to priming different discrete emotions.

Appendix E

Definitions of Eight Discrete Emotions

Table E1. Definition of Eight Discrete Emotions

Discrete emotion	Definition
Surprise	Surprise has one core appraisal, which is whether something is novel or unexpected (Scherer, 2001).
Anticipation	Anticipation occurs in a situation of looking for a purpose; it implies that the situation is predictable (Plutchik, 1984).
Love	Love is conceived of as a blend of joy and acceptance (Plutchik, 1984).
Joy	Joy, according to the Oxford English Dictionary, is a feeling of great happiness.
Anger	Anger is an emotional state that motivates a person to alleviate personal harm attributed to others and is characterized by states of heightened arousal or activation (Yin et al., 2014).
Anxiety	Anxiety is an unpleasant affective state characterized by uneasiness and uncertainty (Smith & Ellsworth, 1985).
Sadness	Sadness arises from loss and helplessness (Lazarus, 1991) and evokes the implicit goal of changing one's circumstances (Lerner et al., 2004).
Disgust	Disgust revolves around the appraisal theme of being too close to an indigestible object or idea (Lazarus, 1991).

Appendix F

Balance Checks

In Table F1, we provide the balance checks on all of the observable characteristics, i.e., gender, age, frequency of watching movies, and the extent to which they consider online reviews before they go to the movies (considering reviews, for short). Then, a Pearson's chi-square test was applied to test whether the assignment was balanced (not significant indicates that we cannot reject the null hypothesis of the balanced assignment). The Pearson's chi-square test is commonly used for testing the statistical independence of categorical variables (e.g., group assignment and gender). We found that all characteristics are balanced ($p > 0.05$) except the frequency of watching movies ($p < 0.01$). To get unconfounded and robust treatment effects, we used a regression approach to explicitly control for all four factors.

Table F1. Balance Checks

	Gender			**Age**			**Frequency of watching movies**			**Considering reviews**		
Group	**Female**	**Male**	**Other**	**21-39 years old**	**40+ years old**	**≤20 years old**	**<once a month**	**Once a month to once a week**	**More than once a week**	**Yes**	**Maybe**	**No**
Positive	31	42	0	52	21	0	25	32	16	55	14	4
Love	19	40	0	34	25	0	43	16	0	47	12	0
Joy	28	32	0	30	29	1	37	22	1	46	12	2
Surprise	17	52	0	34	35	0	42	26	1	54	11	4
Anticipation	35	33	0	37	31	0	59	9	0	48	19	1
Negative	17	34	0	31	20	0	10	31	10	38	11	2
Anger	29	32	1	29	33	0	45	14	3	42	17	3
Anxiety	23	43	0	38	28	0	13	39	14	53	12	1
Disgust	31	34	0	23	42	0	44	17	4	48	16	1
Sadness	23	38	0	29	32	0	41	17	3	46	12	3
p-value of chi-square test ($n = 634$)	0.18			0.07			0.00			0.81		

Appendix G

Variables Measured in the Experiment

Table G1. Variables Measured in the Experiment

Variable	Measurement
Emotion (Yin et al., 2014)	In your opinion, to what extent does each of the following words describe how the reviewer felt when he or she wrote the above review? (Surprise, Anticipation, Love, Joy, Anger, Anxiety, Disgust, Sadness) (1 = not at all, 9 = very much)
Perceived Processing Fluency (Storme et al., 2015)	To what extent do you agree with the following statements when you read the review? • *I have trouble fully understanding the meaning.* • *I get the meaning easily.* • *I understand the message without any problem.* • *I find it is complicated to get the message.* • *I have no difficulty understanding the meaning.*
Perceived Writer Rationality (Xiao et al., 2018)	To what extent do you agree with the following statements when you read the review? • *The writer presented a justification for his or her information.* • *The writer used logic to express his or her viewpoints.* • *The writer explained the reasoning behind his or her review.*
Perceived Helpfulness (Sen & Lerman, 2007)	Using the scale below, how would you describe the above consumer review? • *not at all helpful/very helpful* • *not at all useful/very useful* • *not at all informative/very informative*
Purchase intention (Lee et al., 2014)	• *The likelihood of your watching the movie* • *The probability that you would consider watching this movie* • *Your willingness to watch the movie*

Appendix H

Bonferroni Correction

In statistics, the Bonferroni correction is a method to counteract the multiple comparisons problem, i.e., the risk of false significance increases when considering multiple statistical tests simultaneously. Although the classical Bonferroni correction can mitigate the concern of false significance, it is a conservative method that gives the chance of failure to reject a false null hypothesis, because it ignores potentially valuable information (Simes, 1986; Holm, 1979). Therefore, in what follows, we employ an improved Bonferroni procedure (Simes, 1986).

Formally, let $P_{(1)}, \ldots, P_{(n)}$ be the ordered p-values ($P_{(1)}$ is the smallest value) for testing hypotheses $H_{(1)}, \ldots, H_{(n)}$. $H_{(i)}$ ($i \in \{1, \ldots, n\}$) is rejected provided that $P_{(i)} < i\alpha/n$ and $H_{(1)}, \ldots, H_{(i-1)}$ have all been rejected, where α is the desired overall alpha level and n is the number of hypotheses. In our model, $H_{(i)}$ is the hypothesis that a certain regression coefficient is significantly different from zero and n is the total number of coefficients in 11 regressions in Table 11. After the improved Bonferroni correction, we find that all significant coefficients derived by the ordinary least square estimators remain significant with only two exceptions, i.e., the coefficients of anger and sadness in Column 7 of Table 11. For anger, as we have shown that the effect of anger on PH is completely mediated by perceived writer rationality (PR), this change does not affect our conclusions. For sadness, as its coefficient in Column 8 is smaller than its coefficient in Column 6, we are still confident that PF mediates its effect on PH.

Appendix I

Additional Analysis of Perceived Processing Fluency

We used regression analysis to check the association between PF and three outcome variables that are linked to perceived review helpfulness. The first outcome variable is perceived writer rationality (Xiao et al., 2018). The second is perceived writer effort (PE) (Yin et al., 2014, 2017), i.e., the participants' ratings to the question "In your opinion, how much effort had the reviewer put into writing this review?" The third is the perceived informativeness of the review (PIN), i.e., the participants' ratings to the question "How would you describe the above consumer review, from not at all informative to very informative?" To get robust estimates, we controlled for participants' frequency of movie watching, ages, genders, and whether they consider online reviews before purchasing a movie ticket. Further, we allowed clustered standard error at the group level to account for potentially within-group correlations that may bias the estimates. All coefficients of PF are significantly positive (0.38, 0.21, 0.50, *ps* < 0.01), indicating that higher perceived processing fluency is associated with higher perceived writer rationality, effort, and review informativeness. These results imply that online review readers may perceive the writer as more competent when they find that it is easier to process the information contained in the review. This result may motivate future research to further explore the mechanisms between perceived processing fluency and perceived review helpfulness.